# Analyse de l'articulation entre parole et geste dans un corpus multimodal

MÉMOIRE DE DEA

(Informatique Fondamentale et Applications)

par

Frédéric Landragin

Soutenu le 13 juillet 1999 devant le jury : Jean Berstel, Professeur à l'Université de Marne-La-Vallée
Laurent Romary, Chargé de Recherche CNRS au LORIA

---

Laboratoire Lorrain de Recherche en Informatique et ses Applications – UMR 7503
Université de Marne-La-Vallée – Institut Gaspard Monge

# Table des matières

# Remerciements

Je tiens à exprimer tous mes remerciements à Jean-Marie Pierrel et à Laurent Romary pour m'avoir accueilli dans l'équipe Langue et Dialogue du LORIA, m'ayant ainsi permis de travailler dans un environnement scientifique très profitable.

Je remercie vivement Nadia Bellalem pour m'avoir encadré et apporté tous les éclaircissements nécessaires dans mon travail, ainsi que Susanne Alt, Bertrand Gaiffe, Jean-Luc Husson, Patrice Lopez et Frédéric Wolff, dont les explications et les conseils m'ont été utiles.

Merci aussi aux autres membres de l'équipe : Evelyne, Nadia, Patrice et les autres, ainsi qu'à Armelle, Emma, Sandrine et Yannick, pour leur accueil.

# Introduction générale

### Environnement du mémoire

L'équipe Langue et Dialogue du LORIA comprend une vingtaine de personnes (en comptant les chercheurs doctorants), des informaticiens en majorité et quelques linguistes. Son objectif est de définir des modèles et des techniques permettant de mettre en œuvre des systèmes de dialogue homme-machine finalisés reposant sur une forte composante langagière. Ses directions de recherche sont l'étude des mécanismes fondamentaux de la communication en langue naturelle seule ou accompagnée d'une désignation gestuelle (communication multimodale), la définition d'outils et de méthodes génériques d'études de situations de dialogues, et la réalisation de systèmes de dialogues effectifs.

La désignation gestuelle est un thème de recherche relativement récent et peu exploré. Au LORIA, la thèse de Nadia Bellalem [Bellalem 95], puis la thèse de Frédéric Wolff [Wolff 99] ont porté sur la reconnaissance et l'interprétation contextuelle du geste de désignation, mettant en particulier l'accent sur la nécessité de tenir compte du contexte perceptif visuel lors de l'interprétation. D'autre part, la référence est depuis plusieurs années un thème de recherche central de l'équipe, avec les travaux de Bertrand Gaiffe et d'Anne Reboul, qui se focalisent actuellement sur la définition d'un modèle de représentation mentale des référents [Reboul 98]. Ce modèle inclut, dans la mesure du possible, l'ensemble des informations attachées à un objet et susceptibles d'être activées lors d'une référence à cet objet.

### Sujet du mémoire

Le travail exposé dans ce mémoire est une première étape vers l'intégration du geste de désignation dans la problématique de la référence. À partir d'un corpus multimodal constitué par Frédéric Wolff, nous étudierons l'articulation entre parole et geste de la manière suivante : selon une approche linguistique, nous analyserons les énoncés oraux du corpus en considérant que les gestes de désignation ont été reconnus et interprétés, c'est-à-dire en partant des résultats du travail de Frédéric Wolff. Notre étude consistera ensuite à identifier les problèmes qui se posent lors de l'interprétation des énoncés multimodaux, puis à identifier les informations nécessaires à cette interprétation. Nous établirons alors une classification des mécanismes de référence multimodale, classification qui sera à l'origine d'un modèle d'interprétation. Afin d'élargir la problématique, nous proposerons enfin une synthèse théorique de notre analyse, qui permettra d'aboutir à une classification des mécanismes de désignation et à des possibilités de prolongements de l'étude.

La première partie de ce rapport présente un état de l'art sur la communication homme-machine dans le cadre du dialogue multimodal finalisé, ainsi que sur la référence dans ce contexte. Dans la seconde partie, nous expliciterons notre démarche d'étude et, après une présentation globale du corpus et de l'expérience qui a abouti à sa constitution, nous présenterons les résultats de notre analyse de l'articulation entre parole et geste dans ce corpus, ainsi que notre synthèse théorique.

**Problèmes connexes**

Ce sujet s'inscrit dans plusieurs problématiques étudiées généralement de manière distincte. L'étude de certains aspects de la compréhension dans le dialogue multimodal a déjà été l'objet d'un stage en 1998 au Laboratoire Central de Recherche de THOMSON-CSF [Landragin 98], et ce mémoire de DEA suit la même démarche sur des problèmes relativement proches. Ainsi :

- Le cadre de notre travail est la compréhension de l'entrée multimodale spontanée et non l'étude ergonomique de la multimodalité.
- Nous ne reviendrons pas sur l'interprétation du geste du désignation.
- Nous n'étudierons pas de manière approfondie le problème complexe que constitue la référence, mais nous nous focaliserons sur la référence multimodale dans le dialogue finalisé, sujet sur lequel ne porte qu'un nombre réduit de travaux.
- En conséquence au point précédent, nous n'étudierons pas les possibilités d'adaptation du modèle de représentation mentale de Bertrand Gaiffe et Anne Reboul au cas de la multimodalité : ce travail de plus long terme ne peut pas entrer dans le cadre d'un stage de DEA.

# Première partie

# Problème de la référence dans le dialogue homme-machine multimodal

# Introduction : Le dialogue homme-machine

### Communication homme-machine spontanée

Avec l'essor des technologies de communication, l'ordinateur devient de plus en plus un outil de communication. À l'époque où les seuls utilisateurs étaient des spécialistes, la communication homme-machine se faisait à l'aide de langages spécialisés. Maintenant que les ordinateurs sont utilisés également par des non-spécialistes, la communication homme-machine doit s'adapter : ce n'est plus à l'utilisateur de faire des efforts pour comprendre la machine mais c'est à la machine de faire des efforts pour comprendre l'utilisateur. La meilleure façon de minimiser les efforts de l'utilisateur est d'autoriser les moyens de communication qu'il maîtrise depuis l'enfance : la parole et le geste venant la compléter. Un système de communication homme-machine véritablement naturel devrait, dans la mesure du possible, comprendre l'utilisateur comme le ferait un interlocuteur humain. Il doit donc autoriser des énoncés oraux et gestuels spontanés, effectués sans contraintes. Le problème qui se pose pour de tels systèmes est la compréhension de ces énoncés, qui peuvent être très divers et qui mettent en jeu des mécanismes extrêmement complexes.

### Dialogue homme-machine finalisé

Nous nous plaçons dans le cadre du dialogue homme-machine finalisé, c'est-à-dire du dialogue guidé par l'application, dont le but est de réaliser une tâche. Dans ce type de dialogue, l'utilisateur voit la machine comme un partenaire coopératif : il lui indique les paramètres de la tâche, la machine lui demande éventuellement des précisions et, lorsque tous les paramètres nécessaires sont connus, exécute la tâche. Les dialogues du type *commande de processus* sont des exemples de dialogues finalisés ; citons l'aménagement d'intérieurs, le montage de séquences vidéo, le pilotage d'un robot ou d'une caméra. Pour de telles applications, l'entrée vocale peut être avantageusement complétée par le geste, comme c'est le cas dans la communication homme-homme. Le geste permet en effet de désigner les objets de l'action et de réduire leur description dans l'énoncé oral. Des systèmes de dialogue multimodaux intelligents s'avèrent nécessaires pour traiter l'entrée orale et l'entrée gestuelle.

### Compréhension des énoncés de l'utilisateur

La compréhension d'un énoncé par le système, c'est-à-dire la construction d'une représentation interne de cet énoncé pour déclencher un processus de l'application, est facilitée du fait de l'importance de la tâche. En effet, les connaissances mises en jeu à tous les niveaux sont réduites et permettent d'aborder plus aisément les différents traitements. Ainsi, le lexique et les constructions syntaxiques utilisées au cours d'un dialogue finalisé sont relativement limités, ce qui permet en particulier d'augmenter sensiblement les performances de la reconnaissance de la

parole. De même, la variété de sens des énoncés oraux est limitée par le nombre réduit de commandes possibles. Dans le dialogue uniquement oral, la compréhension des énoncés oraux se fait en tenant compte du contexte applicatif et du contexte discursif (prise en compte des énoncés précédents pour comprendre les ellipses et les anaphores).

Cependant, dans notre cadre du dialogue multimodal, la compréhension d'un énoncé nécessite en plus la prise en compte du contexte lié à la perception visuelle. Ceci peut s'avérer très complexe, d'une part en raison de la variabilité de ce contexte, d'autre part en raison de l'interférence possible entre le contexte perceptif et le contexte discursif.

### Notre démarche

D'une manière générale, notre démarche présente deux aspects :

- Un aspect pragmatique de la référence, ayant conduit d'une part à une analyse théorique des critères de saillance visuelle intervenant lors de la compréhension (dans le cadre du stage à THOMSON-CSF [Landragin 98]), et conduisant d'autre part à une analyse théorique des mécanismes de désignation ainsi que de leurs rapports avec la référence (dans le cadre de ce mémoire).
- Un aspect lié au traitement informatique, conduisant à l'étude des problèmes techniques posés par le traitement automatique des énoncés oraux et gestuels.

Notre état de l'art se décompose ainsi en deux chapitres, le premier sur l'aspect lié au traitement informatique, et le second sur l'aspect lié à la pragmatique de la référence que nous considérons dans le cadre défini par le premier chapitre.

# Chapitre 1

# Communication homme-machine spontanée

## 1.1 Traitement de l'entrée vocale

### 1.1.1 Avantages et inconvénients de la parole

La parole dans la communication homme-machine est incontournable quand les autres canaux de communication sont *saturés* (cas de la communication entre un pilote et son avion) ou *inopérants* (cas des personnes handicapées) [Pierrel 87].

Lorsque ce n'est pas le cas, la parole n'est pas indispensable : ce qui peut être fait avec la parole peut aussi être fait avec des systèmes classiques de menus ou d'icônes. Elle peut cependant être très utile, non seulement parce qu'elle est un moyen naturel de communiquer comme nous l'avons vu dans l'introduction, mais aussi parce qu'elle procure de nombreux avantages :

- Elle est **concise** dans la mesure où un seul énoncé oral peut regrouper plusieurs commandes *identiques* (par exemple : "iconifie *toutes* les fenêtres" au lieu d'iconifier les fenêtres l'une après l'autre) ou même *différentes* (par exemple, dans une application d'aménagement d'intérieurs : "ajoute une chaise en plastique blanc" au lieu de la série de commande : `ajouter(chaise,scène)`, `matériau(chaise,plastique)`, `couleur(chaise,blanc)`.
- Elle est **rapide** *(en plus de la rapidité induite par le point précédent)* dans la mesure où l'utilisateur n'a pas besoin d'explorer les menus et les sous-menus de l'application pour trouver la commande correspondant à ce qu'il veut faire.
- Elle est **confortable** dans la mesure où l'utilisateur prend du recul par rapport à sa machine : d'une part, il n'est pas gêné par l'affichage de ces mêmes menus et sous-menus devenus inutiles ; d'autre part il se place dans une logique de *faire faire* et non de *faire* [Pouteau 94].
- Elle est **efficace** *(en plus de l'efficacité induite par les trois points précédents)* dans la mesure où elle permet à l'utilisateur de se concentrer sur la tâche qu'il doit effectuer tout en relâchant son attention sur les moyens de l'effectuer. Ainsi, l'utilisateur n'a pas à faire alterner son regard de l'écran au clavier ou à la souris [Mathieu 97].

Hors du cadre du dialogue homme-machine finalisé pour lequel ces avantages sont d'importance cruciale, un grand nombre d'applications informatiques seraient beaucoup plus conviviales si elles autorisaient et traitaient l'entrée vocale. Une application de traitement de texte serait par

exemple plus efficace si elle était associée à une machine à dicter. De même, l'accès à des bases de données pourrait se faire par des requêtes vocales (en particulier dans les systèmes d'aide des logiciels ou des matériels complexes). À plus long terme, tous les appareils électroniques grand public, du magnétoscope à la machine à laver, pourraient se piloter par la voix. De plus, l'entrée vocale contribue à l'apparition de nouvelles applications telles que les systèmes d'apprentissage de la prononciation d'une langue ou les systèmes de traduction automatique couplant la reconnaissance vocale, la traduction de texte à texte et la synthèse vocale.

En général et dans le cadre du dialogue homme-machine finalisé, la parole présente cependant un certain nombre d'inconvénients :

- Elle est **bruyante** : le bruit engendré par un utilisateur peut d'une part gêner les personnes proches de lui, d'autre part empêcher la confidentialité des données énoncées.
- Elle ne s'utilise correctement que pour des **processus discrets** et non continus : déplacer un objet par la voix peut s'avérer difficile (par exemple : "un peu plus à droite, encore un peu, non, pas tant").
- Elle peut devenir **contraignante** pour un utilisateur spécialiste qui préfèrera par exemple la saisie de commandes au clavier.
- Elle est **fatigante** : son utilisation prolongée entraîne une certaine fatigue physique.

Ces problèmes ne se posent pas dans l'exemple de la communication entre un pilote et son avion : le pilote est seul ; il utilise la parole de manière ponctuelle pour les processus discrets, les autres canaux de communication étant utilisés simultanément pour les processus continus, ce qui permet une meilleure efficacité. C'est pour ce type de systèmes homme-machine que le traitement automatique de la parole montre tout son intérêt.

### 1.1.2 Reconnaissance automatique de la parole

L'objectif de la reconnaissance de la parole est de reconstituer le message. L'objectif de la compréhension de la parole (que nous présenterons en 1.1.3) est de saisir le sens du message. Alors que les machines à dicter n'intègrent que la reconnaissance, les systèmes de dialogues homme-machine finalisés intègrent la reconnaissance et la compréhension de la parole.

Les avantages de l'entrée vocale cités en 1.1.1 ne sont valables que si la parole est produite sans contrainte, donc en continu (et non mot à mot) et dans un langage naturel (et non dans un langage artificiel). Le traitement de ces énoncés oraux pose des problèmes très complexes (les spécificités lexicales et grammaticales de la langue française parlée seront étudiées en 1.1.4) [Husson 98] [Mathieu 97] [Pierrel 87] :

1. **Problèmes intrinsèques à la parole**, qui ne gênent en rien la communication homme-homme (même des énoncés incompréhensibles sont reconstitués à l'aide du contexte) mais rendent la reconnaissance automatique de la parole très difficile :
   - Sa **continuité** et son imprécision : la parole est la réalisation acoustique d'une séquence discrète de phonèmes qui, non seulement sont parfois difficiles à distinguer les uns des autres, mais de plus peuvent être mal produits et donc mal perçus ; les phonèmes correctement perçus immédiatement avant et après permettent généralement la reconstitution du message.
   - La **variabilité intra-locuteur** : les caractéristiques vocales d'un même locuteur varient dans le temps (phase initiale avant que la voix ne se stabilise, fatigue vocalique

en fin de journée) ou selon les conditions de production de la parole (fatigue, stress) ; un locuteur est même incapable de faire deux fois la même chose.

- La **variabilité inter-locuteur** : chaque locuteur possède des caractéristiques qui lui sont propres, d'une part les caractéristiques de son conduit vocal, d'autre part son accent régional.

- Les phénomènes de **coarticulation**, qui résultent du fait que la réalisation acoustique d'un phonème varie selon les sons précédents et les sons suivants. On observe ainsi :
  - *Au sein d'un même mot* : la disparition du e muet (demande : [dəmãd] → [dmãd]) ; l'assimilation qui correspond à un transfert d'une caractéristique phonétique d'un son sur un son immédiatement voisin (absurde : [absyrd] → [apsyrd]) ; la fusion de deux consonnes identiques avec allongement de la consonne résultante.
  - *À la jonction de deux mots* : les liaisons entre une consonne finale et une voyelle initiale (parfois subtile : "vous êtes innocent" *pas de liaison*, mais "vous êtes innocents" *liaison*) ; les liaisons provoquant des altérations : assourdissement, dénasalisation [divɛ̃] + [ɛspwar] → [divinɛspwar], insertion de phonèmes pour éviter des hiatus [urs] + [blã] → [ursəblã], élision de phonèmes [povr(ə)] + [tip] → [povtip], application des règles d'assimilation ou de fusion [ər(ə)] + [ɛ] + [dəmi] → [ərɛnmi].

- L'effet Lombard ou **compensation des difficultés de perception** que peut rencontrer l'interlocuteur (à cause par exemple du niveau de bruit ambiant), et qui pousse le locuteur à modifier sa production de parole, ce qui se traduit par une réorganisation des mouvements articulatoires et entraîne des variations importantes dans les caractéristiques acoustiques de la parole.

2. **Problèmes liés à l'implantation d'un système de reconnaissance automatique de la parole :**
   - Problèmes liés à l'**acquisition du signal acoustique** : qualité et conditions d'utilisation du dispositif de saisie (micro utilisé à distance fixe, volume de voix ni trop faible ni trop fort, indication du début et de la fin de l'énoncé oral à l'aide de commandes vocales spéciales ou d'un dispositif tel qu'une pédale) ; caractéristiques de l'environnement (bruit de fond dont celui de la machine, conversation des collègues de bureau).

   - Problèmes liés à la **reconstitution de l'énoncé oral** compte tenu des problèmes intrinsèques à la parole : c'est le cœur du problème de la reconnaissance automatique de la parole. Les premiers systèmes utilisaient des techniques de reconnaissance de formes pour comparer chaque mot à reconnaître aux différentes formes de référence stockées. Cette méthode globale s'est révélée insuffisante face à de grands vocabulaires, et une deuxième approche a consisté à segmenter le message en constituants élémentaires *(phonèmes, demi-syllabes, syllabes)*, à identifier ces constituants et à reconstituer la phrase prononcée. Cette méthode analytique a été abandonnée suite aux difficultés d'identification des constituants et aux limites de ses résultats compte tenu de la complexité de sa mise en œuvre. Les systèmes actuels utilisent des méthodes stochastiques : une phase d'apprentissage sur des corpus permet d'obtenir des données statistiques qui sont ensuite utilisées pour classer les hypothèses de reconnaissance. Ainsi, des suites de mots ou de classes syntaxiques peu fréquentes dans le corpus

d'apprentissage seront considérées comme improbables. Dans le cadre du dialogue finalisé pour lequel la variété des énoncés est limitée, la prise en compte de contraintes lexicales, syntaxiques, sémantiques et même pragmatiques permet de supprimer les hypothèses ne correspondant pas au modèle du langage défini pour l'application. Le principal problème qui se pose est le compromis entre la couverture et la précision : un équilibre reste à trouver entre un modèle de langage exhaustif impliquant peu de contraintes (langage plus naturel) et donc une reconnaissance plus difficile, et un modèle de langage réduit impliquant de fortes contraintes (langage moins naturel) et une reconnaissance plus facile. Enfin, tout système intégrant une entrée vocale doit prévoir d'éventuelles erreurs de reconnaissance et doit donc élaborer des stratégies de traitement particulières.

- Problèmes liés aux **besoins *temps réel*** de l'application compte tenu de la nécessité de temps de réaction proches de ceux d'un homme et de la puissance de la machine.

### 1.1.3 Compréhension automatique de la parole dans le dialogue

Comprendre un énoncé, c'est générer une représentation informatique de cet énoncé, représentation qui soit directement utilisable par le système. Ainsi, dans le cadre du dialogue finalisé, il s'agit de retrouver les commandes de l'application à exécuter, c'est-à-dire en pratique les fonctions et leurs paramètres d'appel[1]. C'est ce que l'on appelle *résoudre les références aux actions*. Lorsque l'application manipule des objets (icônes, boutons, objets virtuels, etc) et que les énoncés portent sur ces objets, la compréhension de ces énoncés passe aussi par la *résolution des références aux objets*.

La compréhension de l'entrée vocale dans le dialogue a une particularité que ne présentent pas les autres modalités : elle peut être liée aux énoncés précédents en plus de l'énoncé courant. Deux phénomènes sont en cause :

- **L'ellipse :** une ellipse est la suppression volontaire d'un élément nécessaire à une construction syntaxique complète. Par exemple, dans un dialogue de commande au cours duquel revient souvent le même verbe, on peut observer au bout d'un certain temps une ellipse de ce verbe.

- **L'anaphore :** « Un segment de discours est anaphorique quand il est nécessaire, pour lui donner une interprétation, de se reporter à une autre partie du discours. Les pronoms constituent un ensemble particulier d'anaphore » [Carré 91].

Pour que l'énoncé courant soit correctement traité, il est nécessaire de résoudre les ellipses et les anaphores. La résolution d'une ellipse se fait en cherchant dans le dernier énoncé syntaxiquement complet le composant qui manque à l'énoncé courant. La résolution d'une anaphore est un problème très complexe car non seulement les anaphores peuvent prendre des formes très diverses, mais en plus ce phénomène interagit avec celui de la référence aux objets. Nous y reviendrons dans le chapitre 2.

Dans le cadre du dialogue homme-machine finalisé, la compréhension est liée à la gestion du dialogue, c'est-à-dire au contrôle de son déroulement. Compréhension et gestion du dialogue nécessitent des connaissances que l'on peut répartir en deux catégories [Pierrel 87] :

1. **Connaissances statiques :**
   - ***Le modèle du langage :*** il comprend les composantes lexicale, syntaxique, séman-

1. [Duermael 94] propose de définir une application comme une bibliothèque de fonctions informatiques.

tique et pragmatique qui permettent d'une part de guider la reconnaissance comme nous l'avons vu en 1.1.2, et d'autre part de comprendre un énoncé dans la mesure où le modèle du langage est une représentation du fonctionnement de la langue (la résolution des références s'appuie sur lui).

- ***Le modèle de l'application :*** il définit les objets manipulés ainsi que les relations entre ces objets ; il spécifie en termes de buts et de sous-buts l'accès aux données de l'application.
- ***Le modèle du dialogue :*** il spécifie les stratégies de gestion du dialogue en fournissant une description des diverses situations de dialogue spécifiques à une application et en précisant les liaisons entre ces situations.

**2. Connaissances dynamiques :**

- ***L'univers de l'application :*** il décrit l'état courant des objets de l'application.
- ***L'historique du dialogue*** ou mémoire à long terme : il gère une représentation discursive du dialogue incluant les divers énoncés ainsi que leurs enchaînements (il permet ainsi de résoudre entre autres les ellipses).
- ***Le focus*** ou mémoire à court terme : il contient les résultats de l'analyse du dernier énoncé de l'utilisateur.

Notons qu'un modèle de l'utilisateur peut dans certains systèmes faire partie des connaissances dynamiques [Luzzati 95]. C'est le cas (du moins en théorie) si le système peut s'adapter aux variations des paramètres de la voix du locuteur. C'est le cas également si le système a des capacités d'apprentissage lui permettant de retenir au fur et à mesure du dialogue les préférences langagières du locuteur.

Notons enfin que toutes ces connaissances sont très fortement dépendantes les unes des autres, et qu'il est donc très difficile de les répartir dans différents modules. La définition d'une architecture modulaire et générique aux systèmes de dialogue finalisé est un problème que la multimodalité rend encore plus complexe.

### 1.1.4 Prise en compte des spécificités de la langue parlée

La communication orale en langue naturelle spontanée se fait dans un niveau de langue qui n'est pas celui de l'écrit. L'étude des rapports entre oral et écrit a fait l'objet d'un travail dans le cadre de ce DEA [Landragin 99]. Sans entrer dans les détails, nous noterons qu'il existe des tendances et erreurs grammaticales propres à l'oral. Parmi les constructions syntaxiques utilisées beaucoup plus souvent à l'oral qu'à l'écrit, citons :

- Les phrases emphatiques[2] ou constructions à topiques : "Le camion, il avance" ; "La télé, les vieux, ils la regardent" ; "Jean, sa sœur, je la deteste" *(construction à double topique)* ; "La mer, tu vois de l'eau" *(construction à topique non lié)*[3].
- Les constructions anaphoriques, et en particulier les anaphores (et les cataphores) nulles : "*[il]* Mérite des baffes, ce petit con" ; "*[il est]* Bizarre, ce truc" (exemples de [Lambrecht 96]).
- Les phrases construites sur la tournure "c'est que" : "c'est lui qui" ; "C'est à toi que".
- Les phrases construites sur la tournure "il y a", dont la syntaxe est souvent maladroite.

2. Phrases construites avec reprise d'un groupe du nom ou d'un groupe de l'adjectif par un pronom.
3. Ces deux derniers exemples sont décrits dans [Lambrecht 96].

La langue orale présente d'autre part des structures syntaxiques particulières, liées à sa production et au déroulement du dialogue. Les phénomènes déterminant ces structures peuvent être classés en deux catégories [Lopez 99] :

1. **Phénomènes de bruits et de distorsions**, qui aboutissent à des structures agrammaticales (qui ne seraient en aucun cas admises à l'écrit) :
   - ***Les hésitations, pauses et interjections :*** les hésitations, qui se manifestent par une interjection (mot parasite) ou une pause, permettent au locuteur de se donner le temps de réfléchir à la suite de son énoncé : les interjections comme "euh" ou "hum" meublent le canal de communication afin de préserver l'attention de l'interlocuteur ; les interjections comme "ben" permettent au locuteur de structurer son discours.
   - ***Les répétitions :*** elles permettent au locuteur de se donner le temps de réfléchir à la suite de son énoncé, de se remémorer l'endroit interrompu de l'énoncé, et, d'une certaine manière, de forcer la production de la suite.
   - ***Les précisions :*** une précision est une expression venant enrichir un premier terme choisi rapidement lors de la production de l'énoncé ; elle n'efface pas ce premier terme.
   - ***Les corrections ou reformulations :*** elles diffèrent des précisions en ce qu'elles effacent le premier terme qui doit donc être ignoré ; elles peuvent être introduites par une marque d'hésitation, par une particule négative ("non") ou par une marque d'excuse ("pardon").
   - ***Les reprises :*** une reprise se définit comme une expression cette fois-ci interrompue (parfois au milieu d'un mot), suivie d'une nouvelle version corrigée de cette expression.
   - ***Les effacements ou omissions :*** ils correspondent à un manque involontaire dans une structure syntaxique (par exemple : la chute du discordantiel "ne" des négations).

2. **Phénomènes de fragmentations et d'ellipses**, qui peuvent aussi être utilisés à l'écrit ou qui ont des équivalents à l'écrit :
   - ***Les ellipses :*** comme nous l'avons vu en 1.1.3, il s'agit d'un manque volontaire dans la structure syntaxique, celle-ci pouvant être complétée à partir des structures précédemment employées.
   - ***Les juxtapositions ou parataxes :*** elles désignent la présence dans un même énoncé de plusieurs propositions complètes qui ne sont pas liées par des connecteurs mais par des liens implicites (reformulation, apposition, coordination, énumération, etc).
   - ***Les dislocations :*** elles correspondent à l'incise d'une proposition complète au sein d'une autre proposition, sans connecteur discursif.

Les formalismes sont faits pour l'écrit et n'acceptent a priori aucune des spécificités de la langue orale. Afin de traiter les énoncés oraux comportant des mots parasites, diverses heuristiques ont été proposées : éliminer ces mots parasites dès le niveau acoustique de façon à ce qu'ils n'apparaissent pas dans le treillis d'hypothèses de la reconnaissance ; éliminer ces mots dans le treillis d'hypothèses à l'aide d'un traitement lexical spécifique ; détecter les arrêts de l'analyseur syntaxique et élaborer des stratégies de reprise du traitement au bon endroit dans la phrase.

Le traitement des énoncés oraux comportant bruits, distorsions, fragmentations ou ellipses s'avère plus complexe. Il est à noter tout d'abord que les approches statistiques ne sont pas compatibles avec ces phénomènes : comme il s'agit de ruptures syntaxiques, la probabilité qu'un mot précis suive un autre mot précis est perturbée. Une première approche consiste à intégrer au niveau de la reconnaissance des modèles prosodiques et des modèles syntaxiques locaux, de façon

à repérer les ruptures. Une deuxième approche, qui lui est complémentaire, consiste à étendre un formalisme en lui ajoutant des règles particulières capables de retrouver une bonne structure syntaxique à partir de celle de l'énoncé (ce sont les règles de réparation de [Lopez 99]).

Constituer des formalismes pour la langue orale est une solution irréalisable. En effet, un formalisme sert par définition à décrire des constructions variées à l'aide d'un nombre réduit d'opérateurs. Or la langue orale comprend tellement de constructions possibles (par exemple une répétition peut avoir lieu à tout moment dans la phrase) que la définition des opérateurs en devient impossible. De plus, un formalisme acceptant toutes les constructions de l'oral serait tellement lâche qu'il finirait par accepter n'importe quelle construction. Prendre un formalisme fait pour l'écrit et l'étendre à certains phénomènes de l'oral semble être une solution raisonnable [Lopez 99].

## 1.2 Traitement de l'entrée gestuelle

### 1.2.1 Gestes co-verbaux

La communication orale ne met pas en jeu que la parole. [Kerbrat-Orecchioni 96] distingue :

- **Le matériau verbal** qui relève de la langue : informations phonologiques, lexicales, syntaxiques, sémantiques.
- **Le matériau paraverbal** (prosodique et vocal) : intonations, pauses, intensité articulatoire, débit, prononciation, caractéristiques de la voix.
- **Le matériau non verbal** transmis par le canal visuel : l'apparence physique des participants, les cinétiques lents (postures, attitudes) et les cinétiques rapides (regards, gestes).

Le geste intervient donc comme un complément naturel de la parole qui peut accroître son efficacité et permettre à l'interlocuteur de mieux comprendre le message. Nous l'étudierons selon cette approche. Cela n'empêche cependant pas le geste de porter une signification qui lui est propre, ni d'avoir une certaine autonomie. Deux classes de gestes peuvent être distinguées selon leur autonomie par rapport à l'énoncé oral [Braffort 96] :

1. **Les gestes symboliques ou emblématiques**, qui sont indépendants de l'énoncé oral et qui peuvent accompagner ou remplacer tout ou une partie de cet énoncé (par exemple le geste de salut). Ce type de geste est relativement courant dans la communication homme-homme, bien que limité à quelques situations précises[4]. Dans la communication homme-machine, ce sont les gestes de commande : geste en forme de croix pour "supprimer", geste en forme de flèche pour "déplacer", etc. Ces gestes ne sont pas naturels dans le sens que leur apprentissage est nécessaire pour une bonne utilisation de l'application. Dans un système de dialogue acceptant l'entrée gestuelle spontanée, nous n'en trouverons que très rarement, et ils seront alors toujours accompagnés d'un énoncé oral explicite.

2. **Les gestes illustrateurs**, qui dépendent de l'énoncé oral dans le sens que la compréhension du message de l'utilisateur nécessite la combinaison du contenu du message oral et de celui du message gestuel. On les appelle aussi *gestes co-verbaux*. On distingue :
   - Les gestes *déictiques* : ils désignent un objet ou un lieu référencé simultanément dans l'énoncé oral.

4. On les trouve également dans les langages gestuels comme la LSF (Langue des Signes Française).

- Les gestes *iconiques* : ils représentent un objet, une action ou un événement référencé simultanément dans l'énoncé oral (par exemple : "J'ai pêché un poisson grand comme ça" + geste d'écartement des deux mains).
- Les gestes *métaphoriques* : ils illustrent un concept abstrait (par exemple : "Il veut s'emparer de nos idées" + geste mimant l'action de saisir un objet).
- Les gestes de *battement* : ils marquent le rythme du discours.

Les gestes illustrateurs sont naturels et se rencontrent fréquemment dans la communication homme-homme. Il n'en est pas de même dans la communication homme-machine : autant les gestes déictiques sont fréquents dès qu'il est question d'objets ou de lieux, autant les autres types de gestes illustrateurs sont rares, du fait de leur caractère conversationnel qui va à l'encontre de la focalisation à la résolution de la tâche précise de l'application. On peut néanmoins trouver quelques gestes iconiques (par exemple : "Agrandir la fenêtre comme cela" + geste en forme de rectangle ou de droite indiquant la largeur voulue).

### 1.2.2 Dispositifs gestuels

Dans la communication homme-homme, le geste effectué à l'aide de la main combine trois paramètres qui sont : l'*emplacement*, l'*orientation* et la *configuration* de la main, ces trois paramètres pouvant varier dans le temps selon les mouvements effectués. Dans une communication homme-machine véritablement naturelle, ces trois paramètres devraient être captés et traités en temps réel. Cependant, le caractère naturel du geste dépend avant tout du dispositif utilisé. Selon le rapport entre l'utilisateur et le dispositif, [Bellalem 95] distingue :

- **Les dispositifs externes :** ils permettent théoriquement la capture des trois paramètres du geste sans aucune instrumentation pour l'utilisateur. Il s'agit des systèmes de caméras couplées à des traitements du type *reconnaissance de formes*. Le principal avantage de ces dispositifs est qu'ils ne gênent pas l'utilisateur qui va pouvoir effectuer des gestes parfaitement naturels, exactement comme dans la communication homme-homme (l'utilisateur pourra même laisser libre cours à des gestes de battement). Parmi les inconvénients de ces dispositifs, la complexité du traitement informatique est un problème extrêmement complexe, difficile à concilier avec les besoins *temps réel* des applications.

- **Les dispositifs à immersion :** ils permettent d'une part la capture de l'emplacement et de l'orientation de la main grâce à un dispositif spécifique *(capteur Polhemus)* qui vient se greffer sur la main de l'utilisateur, d'autre part la capture de sa configuration grâce à une instrumentation venant elle aussi se greffer sur la main. Il s'agit des gants numériques qui, selon les modèles, mesurent les angles de chaque articulation à l'aide d'une structure métallique exosquelettique, de jauges de contraintes, de capteurs à encre conductrice ou même de fibres optiques. Plus les mesures sont précises, plus l'instrumentation est contraignante et donc moins les gestes sont naturels. Compte tenu de cet équilibre à trouver, ces dispositifs présentent une complexité de traitement non négligeable.

- **Les dispositifs de vis-à-vis :** ils ne permettent pas la capture du geste mais mesurent l'action de l'utilisateur sur un objet, qui peut être une souris, un *trackball*, un écran tactile, etc. On assimile cet objet au dispositif. Généralement, le mouvement ne peut se faire que sur un plan (en *2D*), bien qu'il existe des dispositifs *3D*. Ces derniers mesurent l'intensité de pression de l'utilisateur et sont plus faits pour calculer l'intensité d'un geste *2D* que pour calculer une trajectoire *3D*. L'avantage de ces dispositifs est la relative facilité de leur traitement. Leur principal inconvénient est qu'ils placent un intermédiaire entre l'utilisateur

et la machine, ce qui nuit fortement au caractère naturel des gestes. Notons que chaque dispositif permet des gestes plus ou moins naturels ; par exemple, tracer un cercle est plus aisé avec un écran tactile qu'avec une souris.

Les dispositifs les plus utilisés actuellement sont bien entendu les dispositifs de vis-à-vis *2D*. Cela ne doit pas surprendre dans la mesure où la perception de l'application se fait toujours à l'aide d'un écran (*2D*) et, dans une moindre mesure, du fait que la majorité des applications informatiques sont encore en *2D* [5].

D'autre part, nous avons étudié les dispositifs selon notre approche du geste venant en complément de la parole pour émettre de l'information, c'est-à-dire selon sa fonction **sémiotique** (*faire savoir*) telle que définie par [Cadoz 94]. Or le geste dans la communication homme-homme et dans certaines applications homme-machine peut avoir une fonction **épistémique** (*connaître* par le sens du toucher) ou une fonction **ergotique** (*agir* sur l'environnement). Dans ces deux derniers cas, le dispositif gestuel devrait permettre le retour de force. Or seuls quelques dispositifs à immersion et quelques dispositifs de vis-à-vis très particuliers le permettent.

Dans l'état actuel de la technologie, le choix du dispositif gestuel est donc une contrainte de l'application et ne permet pas une entrée gestuelle vraiment spontanée. L'utilisation du geste en est limitée, et c'est pour cette raison que nous ne nous intéressons qu'aux gestes déictiques, comme nous l'avons dit en 1.2.1.

### 1.2.3 Analyse structurelle

Dans cette partie et dans la suite de ce rapport, nous considérons le geste déictique (ou geste de désignation) capté à l'aide d'un dispositif de vis-à-vis *2D* tel qu'un écran tactile, pour des applications *2D*.

Le signal provenant du dispositif gestuel a beaucoup de points communs avec celui provenant d'un micro pour l'entrée vocale. Le problème consistant à analyser un flux continu de données afin d'en extraire des informations précises et structurées est d'ailleurs le même. Ainsi, nous retrouvons exactement les mêmes problèmes de reconnaissance automatique que ceux vus en 1.1.2 :

1. **Problèmes intrinsèques au geste :**
   - Sa **continuité** et son imprécision : le geste dans les conditions que nous avons fixées correspond à une trajectoire, c'est-à-dire à un échantillon de points. Cette trajectoire peut être réduite à un point, mais peut aussi prendre la forme d'un gribouillage ou d'un entourage. Selon la vitesse à laquelle elle est produite, elle est plus ou moins précise, et donc elle cible plus ou moins bien son objectif, par exemple les objets ou le lieu à désigner.
   - La **variabilité intra-utilisateur** : beaucoup plus que la parole, le geste présente une grande variabilité due au contexte (par exemple selon l'objet à désigner). Ainsi, un simple geste d'entourage d'un objet peut prendre des milliers de formes possibles si l'entourage suit les contours de l'objet. D'autre part, la production d'un même geste peut varier dans le temps chez un même utilisateur : la répétition d'une même forme entraîne une simplification de cette forme ; la fatigue ou le stress peuvent entraîner une plus grande imprécision ou une moins bonne fluidité.

5. Il faut en effet distinguer *dispositif de vision*, *dispositif gestuel* et *type d'application* : bien qu'un dispositif de vision *3D* ne se conçoit que dans des applications *3D*, il n'en est pas de même pour le geste. [Landragin 98] présente les problèmes liés à l'utilisation et au traitement du geste effectué à l'aide d'un dispositif *2D* dans une application *3D*.

- La **variabilité inter-utilisateur** : parmi les formes de trajectoire ou dans la manière de cibler l'objectif, chaque utilisateur a ses préférences.
- Les phénomènes de **coarticulation** : lorsque plusieurs gestes s'enchaînent, le début d'un geste peut être modifié par la fin du geste précédent (et inversement).
- Le phénomène de **compensation** qui pousse l'utilisateur à corriger sa trajectoire au cours de sa production. C'est le cas par exemple avec un écran tactile qui fait apparaître la trajectoire du geste : si l'imprécision due au dispositif entraîne un décalage entre la position du doigt ou du stylet et celle des pixels de la trajectoire, l'utilisateur va essayer de compenser en décalant son geste.

**2. Problèmes liés à l'implantation :**

- Problèmes liés à l'**acquisition du signal** : nous les avons explicité en 1.2.2.
- Problèmes liés à la **reconnaissance** compte tenu des problèmes intrinsèques au geste : comme pour la reconnaissance de la parole, il s'agit ici aussi du cœur du problème. Les deux approches classiques, *globale* et *analytique*, sont a priori envisageables. [Bellalem 95] exclue l'approche globale car elle consisterait à identifier et stocker toutes les formes possibles de trajectoires, ce qui s'avère impossible compte tenu de leur grande variabilité. L'approche analytique consiste alors à segmenter le signal afin d'en extraire les parties significatives (par exemple un point d'arrêt dans un entourage). Selon le modèle de [Bellalem 95], une partie significative se caractérise par une singularité, c'est-à-dire une rupture d'homogénéité pour une des propriétés de la trajectoire, les propriétés fondamentales étant la courbure et la vitesse. L'*analyse structurelle* du geste consiste donc à :
  - repérer les singularités ;
  - modéliser la trajectoire avec des courbes *B-splines* ;
  - repérer à partir de cette modélisation d'autres particularités telles que les points d'intersection ;
  - décrire à partir de tous ces éléments le geste en termes de formes et de singularités ;
  - interpréter le geste dans les contextes langagier et spatial, ce qui permet d'éliminer les hypothèses de désignations non valides et de déterminer les objets candidats à la désignation.
- Problèmes liés aux **besoins *temps réel*** de l'application compte tenu de la puissance de la machine.

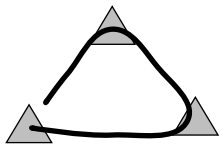
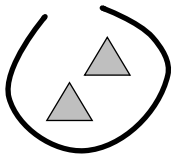
*Remarque : Dans le cas des dispositifs 3D, on observe une phase initiale (correspondant par exemple à l'approche de la main) avant le geste proprement dit, ainsi qu'une phase finale (correspondant par exemple au recul de la main). Le premier objectif de la segmentation est alors de repérer les phases initiales et finales pour ne pas en tenir compte dans la suite du traitement, ce qui peut s'avérer difficile.*

### 1.2.4 Analyse contextuelle

Alors que l'analyse structurelle donne une interprétation du geste à partir de sa trajectoire, l'analyse contextuelle donne une interprétation du geste tenant compte également de l'environnement, c'est-à-dire du contexte perceptif visuel à l'instant de la désignation. Ainsi, dans notre

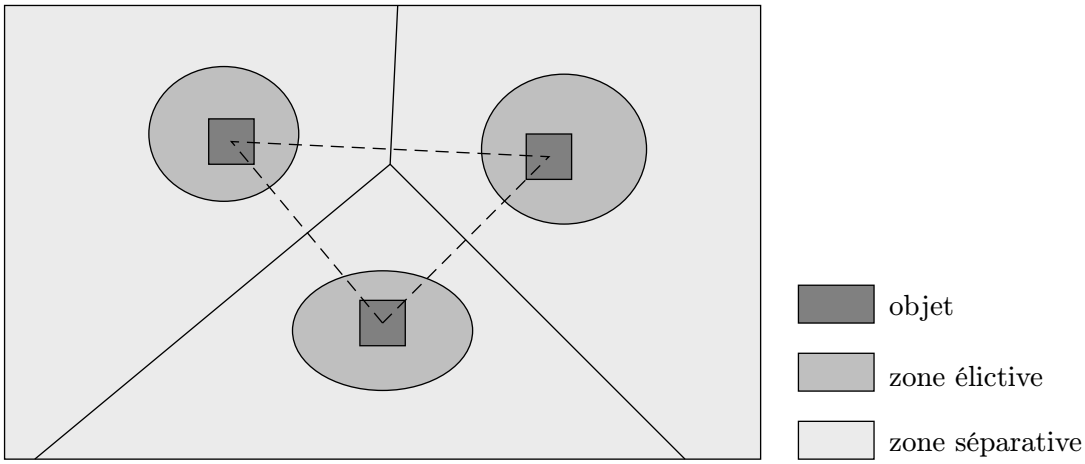
cadre du geste sur écran tactile, la disposition des objets par rapport à la trajectoire du geste va intervenir dans l'interprétation de celui-ci.

Les principaux travaux réalisés sur ce problème sont ceux de [Wolff 99]. Selon une approche psychologique, Frédéric Wolff explicite la notion de *groupement perceptif* : les objets affichés sur l'écran se partitionnent en groupements selon deux critères principaux, tout d'abord la proximité, ensuite la similarité. L'utilisateur perçoit ces groupements, et les gestes qu'il va produire vont en dépendre. Quatre catégories de gestes sont distinguées :

| Pointage | Ciblage | Entourage | Gribouillage |
|---|---|---|---|
|  |  |  |  |

D'autre part, Frédéric Wolff montre que la scène affichée sur l'écran peut se partitionner en plusieurs zones :

- **Les zones élictives :** à chaque objet et chaque groupement est liée une zone de sélection ou *zone élictive* qui le recouvre. Un geste dont la trajectoire reste dans la zone élictive d'un objet ou d'un groupement aura l'intention de désigner cet objet ou ce groupement. C'est typiquement le cas du pointage. La zone élictive s'étend un peu au-delà du gabarit de l'objet ou du groupement, afin de pallier l'imprécision des gestes.
- **La zone séparative :** le reste de la scène, c'est-à-dire le fond qui ne peut correspondre à la sélection d'aucun objet, constitue la *zone séparative*. Un geste dont la trajectoire reste dans cette zone aura ainsi l'intention de séparer certains objets d'autres objets. C'est le cas de l'entourage.

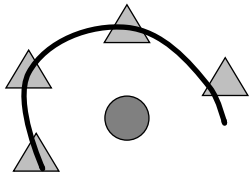
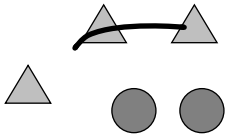



L'intention d'un geste n'est cependant pas toujours aussi claire. Ainsi, un point important qui ressort des travaux de Frédéric Wolff est l'ambiguïté du geste :

- **L'ambiguïté de forme :** une même forme de trajectoire peut correspondre à plusieurs intentions de désignation différentes. Ainsi, un geste ponctuel (pointage) peut désigner

un seul objet ou un groupe d'objets. Si l'on ne tient pas compte de la disposition des objets proches du pointage, on ne peut pas interpréter le geste. De même, un geste de ciblage passant par les centres de gravité de plusieurs objets formant un cercle a une forme circulaire. Si un objet particulier se trouve au milieu du cercle, le geste peut aussi correspondre à l'entourage de cet objet et est donc ambigu.

- **L'ambiguïté de portée :** deux gestes ayant la même forme et correspondant à la même intention de désignation (par exemple un ciblage) peuvent désigner un nombre d'objets différent. C'est le cas lorsque la trajectoire du geste passe par deux objets, un troisième objet du même groupe perceptif se trouvant juste à côté : une première interprétation aboutit à la désignation de deux objets, une autre interprétation aboutit à la désignation du groupe perceptif complet, donc de trois objets.

| Ambiguïté de forme | Ambiguïté de portée |
|---|---|
|  |  |

Enfin, le parallèle entre la compréhension de la parole et la compréhension du geste peut aller plus loin que ce que nous avons vu en 1.2.3 : on peut considérer que les différents types de singularités constituent un lexique, que leurs combinaisons définissent une syntaxe, que l'interprétation de la trajectoire correspond à la sémantique du geste, et que l'interprétation contextuelle correspond à une certaine pragmatique du geste.

## 1.3 Traitement de l'entrée multimodale

### 1.3.1 Caractérisation de la multimodalité

Une interface multimodale combine plusieurs modalités d'entrée et de sortie. Les interfaces classiques combinant le clavier et la souris sont donc multimodales, de même que les interfaces combinant : reconnaissance de la parole ; reconnaissance du geste et de l'écriture ; capture de la direction du regard et de certaines caractéristiques du visage ; et même lecture sur les lèvres de l'utilisateur pour augmenter les performances de la reconnaissance de la parole [Waibel 95]. Dans de telles interfaces, la combinaison des modalités est soumise à des contraintes. [Coutaz 91] puis [Brison 97] proposent des classifications des interfaces multimodales pour décrire l'ensemble des possibilités offertes au concepteur. [Brison 97] part ainsi de trois paramètres qui sont :

- **La production des énoncés :** *séquentielle* ou *parallèle*, un énoncé pouvant utiliser plusieurs modalités. En considérant qu'un énoncé porte sur une tâche particulière, une interface parallèle permet à l'utilisateur de faire plusieurs actions à la fois. C'est le cas par exemple du pilote de chasse qui dirige son avion par le geste tout en commandant ses armes (*visée* grâce à un système de capture de la direction du regard ou plus simplement de l'orientation du casque, et *tir* grâce à la commande vocale "feu !").

- **Le nombre de modalités par énoncé :** *un* ou *plusieurs* (en reprenant l'exemple précédent, la commande des armes fait l'objet d'un énoncé multimodal).

- **L'usage des modalités pour un énoncé :** *exclusif* (si par exemple on considère que l'on tire après avoir visé) ou *simultané* (si l'on considère que l'on est toujours en train de viser au moment de tirer).

| *Type de modalité* | *Production des énoncés* | *Nombre de modalités par énoncé* | *Usage des modalités pour un énoncé* |
|---|---|---|---|
| exclusive | séquentielle | un | – |
| alternée | séquentielle | plusieurs | exclusif |
| synergique | séquentielle | plusieurs | simultané |
| parallèle exclusive | parallèle | un | – |
| parallèle alternée | parallèle | plusieurs | exclusif |
| parallèle synergique | parallèle | plusieurs | simultané |

Contrairement aux interfaces multimodales du type *exclusif*, les interfaces du type *alterné* ou *synergique* posent des problèmes de relations entre modalités ainsi que des problèmes de fusion des informations transmises par ces modalités. Ces problèmes se posent en particulier pour les interfaces permettant les entrées vocale et gestuelle.

D'autre part, chaque modalité a ses propres caractéristiques pouvant perturber les relations entre modalités ainsi que la fusion de leurs informations. Les *temps de réponse* des modalités peuvent par exemple être différents, ce qui pose des problèmes lors de la gestion et de l'interprétation des événements multimodaux. [Mignot 95] distingue de plus deux *points de vue* possibles pour une même modalité : un point de vue *discret* correspondant à un effet ponctuel de la modalité, ou un point de vue *continu* correspondant à un effet duratif de la modalité. Lors de la combinaison des modalités se pose ainsi le problème de la combinaison de ces points de vue :

- **Combinaison de deux points de vue continus :** l'intérêt est de simuler une indépendance entre les deux modes afin de permettre à l'utilisateur le développement de ses capacités de coordination. L'exemple que donne [Mignot 95] est celui d'un système de pilotage d'un tank : à l'aide de deux manettes indépendantes, contrôlant chacune la vitesse de rotation d'une chenille, l'utilisateur peut faire avancer et tourner le char. Cela pose le problème consistant à confronter en *temps réel* les informations provenant de chacune des modalités.

- **Combinaison d'un point de vue continu et d'un point de vue discret :** dans l'exemple d'une interface capable de combiner un geste de dessin avec une commande vocale, le geste peut être utilisé pour tracer une forme continue et la commande vocale pour indiquer les changements d'épaisseur du trait. Cela pose le problème consistant à faire intervenir l'effet d'une modalité au cours de l'effet de l'autre modalité.

- **Combinaison de deux points de vue discrets :** l'aspect *temps réel* n'intervient pas lors de la confrontation des informations provenant des deux modalités.

La commande vocale est un exemple de point de vue discret car son effet est théoriquement immédiat (une fois l'énoncé terminé). Il n'en est pas de même du *geste de commande* : le *drag and drop* à la souris est un exemple de geste dont l'effet, qui a lieu en même temps que le geste puisque l'icône déplacée suit le curseur, prend un certain intervalle de temps. Le geste de désignation a par contre un effet discret : c'est le résultat de la désignation qui ne peut être calculé qu'une fois le geste terminé, exactement comme pour la commande vocale. Le problème de la combinaison d'un point de vue continu et d'un point de vue discret ne se pose donc pas pour l'association de la commande vocale et du geste de désignation.

Dans la suite de ce rapport, nous désignerons par *interface multimodale* une interface bimodale combinant la voix et le geste de désignation, ce qui correspond, comme nous l'avons montré dans les sections précédentes, à une communication naturelle entre l'utilisateur et sa machine. La production simultanée de la parole et du geste étant autorisée, nous nous plaçons dans le cadre des interfaces multimodales *synergiques*, avec une approche d'interprétation d'une modalité par rapport à l'autre.

### 1.3.2 Association de la commande vocale et du geste de désignation

Dans notre cadre du dialogue homme-machine finalisé, l'énoncé oral contient la description de l'action que l'utilisateur veut faire exécuter au système, ainsi que les références aux arguments de cette action, qui peuvent être des *objets* de la scène affichée sur l'écran ou des *lieux* dans cette scène, et qui peuvent être désignés par le geste :

- **Désignation d'objets :** lorsqu'un objet est un argument de l'action à effectuer, l'utilisateur peut décrire l'objet dans l'énoncé oral et/ou le désigner par un geste *(accès individuel)* ; lorsque plusieurs objets sont des arguments de l'action, l'utilisateur peut soit désigner chaque objet l'un après l'autre *(désignation multiple, c'est-à-dire série d'accès individuels)*, soit désigner les objets par un seul geste s'ils forment un groupe perceptif *(accès pluriel)*.

- **Désignation de lieux :** comme pour la désignation directe d'objets, l'utilisateur peut décrire le lieu dans l'énoncé oral et/ou le désigner par un geste ; la désignation multiple de lieux et l'accès pluriel sont également possibles, bien que peu fréquents. Notons également que la désignation d'un lieu pose le problème de l'étendue de ce lieu : dans "mettre de la moquette ici" + geste [Romary 93], et dans "planter un clou ici" + geste, le geste peut être le même (par exemple un pointage) et le lieu désigné de taille très différente. Ce problème rejoint de celui de l'ambiguïté de portée du geste que nous avons vu en 1.2.4.

Les informations données par l'énoncé oral et celles données par l'énoncé gestuel se combinent selon les relations suivantes [Guyomard 95] :

- **Substitution :** dans certains cas, l'utilisateur peut substituer une expression verbale par un geste. L'exemple donné dans [Guyomard 95] est le suivant : "Où sont les campings à . . . ?" + geste sur Morestel (pointage sur une carte géographique). Cette substitution s'explique ici par l'incertitude du locuteur sur la prononciation du nom propre.

- **Complémentarité :** les deux modalités apportent chacune une partie de l'information utile à la compréhension.

- **Redondance :** certaines informations apportées par une modalité sont également apportées par l'autre. Dans certains cas, toute l'information apportée par une modalité est redondante : lorsque par exemple l'expression verbale suffit à la compréhension, le geste peut être considéré comme inutile.

Les problèmes apparaissant lors de la compréhension des énoncés multimodaux sont liés aux références, c'est-à-dire lorsque l'utilisateur désigne par la voix et/ou le geste les objets ou les lieux de l'application. Le traitement des références est un problème complexe qui nécessite une analyse linguistique de l'énoncé oral, comme nous l'avons déjà évoqué en 1.1.3 et comme nous allons le voir en détail dans le chapitre suivant.

# Chapitre 2

# Référence aux objets

## 2.1 Problème de la référence

### 2.1.1 Définitions et approche

Résoudre une référence du discours, c'est faire le lien entre les mots utilisés et les objets de l'environnement, c'est-à-dire les objets du monde réel dans le cas du dialogue homme-homme et les objets de l'application dans le cas du dialogue homme-machine. Résoudre une référence multimodale, c'est faire le lien entre d'un côté les mots et les gestes, de l'autre côté les objets de l'environnement. Nous considérons le geste de désignation comme un support de l'énoncé oral, donc nous ne parlerons pas de référence purement gestuelle.

Nous avons vu en 1.1.3 qu'un énoncé oral pouvait contenir une anaphore, c'est-à-dire deux références successives à un même objet de l'environnement. Un énoncé multimodal contient également deux références à un même objet : une dans l'énoncé oral et l'autre dans le geste. Deux références à un même objet constituent une *co-référence* et on distingue donc (à la suite de [Gaiffe 92]) :

- **co-référence intra-mode asynchrone :** deux références successives au même objet, à l'intérieur d'une même modalité. C'est le cas des anaphores, comme par exemple : "Enlève *les chaises bleues*. Je n'aime pas *ces meubles*."
- **co-référence inter-mode synchrone :** deux références simultanées au même objet, une par modalité[6]. Par exemple : "déplacer *ce meuble*" + geste de désignation en même temps que "*ce meuble*".
- **co-référence inter-mode asynchrone :** deux références successives au même objet, une par modalité. En reprenant le même exemple : "déplacer *ce meuble*" + geste de désignation après la fin de l'énoncé oral.

Le terme de *référence multimodale* regroupe les deux types de co-références inter-modes. Nous avons évoqué en 1.1.3 le problème de l'interférence entre anaphore et référence multimodale. Les exemples donnés ci-dessus le montrent : ils contiennent tous les deux un groupe nominal démonstratif, or le premier est une *reprise anaphorique* et le deuxième est associé à un geste pour effectuer la *désignation directe d'un objet*. A priori, seule la présence d'un geste permet à la machine de distinguer les deux mécanismes.

6. Ce type de co-référence n'est pas possible dans une interface multimodale de type *alternée* (cf. section 1.3.1).

L'étude des références multimodales doit en effet partir des possibilités de la langue : du fait que nous considérons le geste comme un support de l'énoncé oral, c'est de l'énoncé oral et de ses caractéristiques linguistiques qu'il faut partir, non seulement pour identifier les types de références (multimodales ou non), mais aussi pour traiter ces références. [Gaiffe 94] distingue ainsi :

- **La référence directe :** l'expression délivre directement le référent, que celui-ci soit générique (par exemple : "Patrice aime *les desserts*") ou spécifique (par exemple : "enlève *la chaise bleue*", l'environnement comportant une seule chaise bleue).
- **La référence démonstrative :** il s'agit de la référence multimodale telle que nous l'avons définie.
- **La référence anaphorique :** il s'agit de la co-référence asynchrone dans le discours.
- **La référence déictique :** l'expression fait référence au contexte d'énonciation spatial ou temporel, comme par exemple dans : "*Aujourd'hui*, il fait beau *ici*."
- **La référence indirecte :** un référent intermédiaire est utilisé pour mentionner un autre référent, comme par exemple dans : "Je viendrai par *le 13h17*."

Une étude plus poussée des types de groupes nominaux et de pronoms s'avère nécessaire. C'est l'objet de la section suivante, dans laquelle nous étudierons les possibilités de désignation gestuelle pour chaque type d'expression référentielle.

### 2.1.2 Étude des groupes nominaux

Nous avons vu que le groupe nominal démonstratif pouvait correspondre à plusieurs mécanismes. Il en est de même du groupe nominal défini dans les exemples de la section précédente : "*les desserts*", "*la chaise bleue*" et "*le 13h17*" correspondent à trois mécanismes différents. Le but de la classification suivante, obtenue à partir de celle de [Gaiffe 92], est d'identifier les mécanismes possibles pour chaque type d'expression, en se focalisant sur les possibilités de désignation gestuelle :

1. **Les groupes nominaux définis :**
   Leur première fonction est la *référence directe*. Dans l'exemple (1), les informations contenues dans le groupe nominal, ici le nom et l'adjectif, permettent de filtrer les objets de l'environnement.
   La deuxième fonction du groupe nominal défini est la *référence anaphorique*, comme le montre l'exemple (2). L'anaphore peut être *associative*, c'est-à-dire que la deuxième référence est associée à la première par une relation de composition, comme dans l'exemple (3).
   Dans les deux cas, un geste de désignation peut être associé au groupe nominal. Généralement, l'information qu'il apporte est redondante.

   (1) *Déplace* la chaise bleue.
   (2) *Déplace la chaise bleue et le bureau vert. Enlève* la chaise.
   (3) *J'ai acheté un stylo mais* la plume *est cassée.*

**2. Les groupes nominaux démonstratifs :**
Ils semblent avant tout dédiés à l'association avec le geste de désignation pour une fonction de *référence démonstrative*.
Pourtant, la reprise *anaphorique* par le groupe nominal démonstratif existe, comme le montre l'exemple (4). L'anaphore peut se combiner avec une *hyperonymie*[7], comme dans l'exemple (5). Dans les deux cas, le geste de désignation est impossible car il inciterait à considérer un nouvel objet et romprait la relation anaphorique.

(4) *Déplace la chaise bleue. Supprime cette chaise.*
(5) *Déplace le bureau vert. Supprime ce meuble.*

**3. Les pronoms personnels représentants :**
La fonction principale du pronom personnel complément (ce que [Gaiffe 92] nomme *pronom défini*) est l'*anaphore*, comme dans l'exemple (6).
Se pose alors la question de l'antécédent repris par le pronom : est-ce l'information contenue dans le groupe nominal antécédent ou l'objet lui-même ? Reprendre l'information du groupe nominal antécédent nécessite de procéder à un deuxième calcul de référent, alors que reprendre l'objet permet d'utiliser directement le résultat du premier calcul de référent. [Gaiffe 92] montre que ces deux méthodes ne sont pas satisfaisantes. En effet, dans l'exemple (7), la reprise des informations du groupe nominal antécédent donne : "*la*" = "*la chaise verte*" qui justement n'est plus verte. Et dans l'exemple (8), la reprise de l'objet est impossible puisque l'objet n'existe plus après la première référence. Or, dans le dialogue homme-machine finalisé, ces deux exemples sont fréquents et doivent être traités correctement. Une solution consiste à reprendre la représentation sémantique, c'est-à-dire en particulier la structure logique du groupe nominal (par exemple s'il contient une coordination). La structure du groupe nominal est en effet importante, comme le montre l'exemple (9) qui présente une anaphore sous forme d'ensemble.
L'anaphore peut aussi être *divergente*, c'est-à-dire que le pronom ne réfère pas au même objet, comme le montre l'exemple (10) où le pronom élidé réfère à un autre exemplaire du livre antécédent. Le *changement d'état*, illustré par l'exemple (11), est un cas d'anaphore assez proche du précédent.
Enfin, le pronom anaphorique peut impliquer un *glissement au générique*, comme le montre l'exemple (12).
Dans tous les cas, aucun geste de désignation n'est possible. Dans les exemples cités, la présence d'un geste rendrait même l'énoncé multimodal incompréhensible. Il semble que le geste ne peut pas s'associer à une reprise anaphorique.

(6) *Déplace la chaise bleue. Supprime-la.*
(7) *Peints la chaise verte en bleu. Déplace-la.*
(8) *Supprime le grand bureau. Remplace-le par une table basse.*
(9) *Déplace la chaise verte. Déplace la chaise rouge. Mets-les à côté de la table.*
(10) *Ne lui achète pas ce livre. Il l'a déjà.*
(11) *On a rasé la chevelure de Samson, mais elle a repoussé.*
(12) *J'ai acheté une Toyota parce qu'elles sont robustes.*

7. *Meuble* est un hyperonyme de *chaise* ; *table* et *chaise* sont des hyponymes de *meuble*.

4. **Les pronoms démonstratifs :**
   Ils combinent une *référence démonstrative* et une *anaphore* : ils sont associés à un geste pour désigner un nouvel objet ayant les caractéristiques d'un objet référé dans la partie précédente du discours. Dans l'exemple (13), "*celle-ci*" désigne une autre "*chaise bleue*". Contrairement au pronom défini, le traitement du pronom démonstratif se fait en reprenant les informations indiquées dans le groupe nominal correspondant à la première référence. Le problème se ramène alors au traitement de la référence démonstrative.

   (13) *Déplace cette chaise bleue. Supprime celle-ci.*

La diversité des exemples cités montre la richesse de la langue naturelle et la difficulté de son traitement. En effet, autoriser l'entrée vocale spontanée revient à autoriser tous les types de groupes nominaux et de pronoms. Or, à partir du moment où par exemple les pronoms personnels représentants sont autorisés, tous les mécanismes liés à l'emploi de ces pronoms doivent être traités. Dans le cadre du dialogue finalisé, certaines figures de style ne seront jamais utilisées car les situations pouvant amener à leur utilisation ne se présenterons jamais. Avant de négliger le traitement de ces figures de style, il est néanmoins important d'identifier précisément toutes les situations pouvant se présenter dans l'application.

Dans les systèmes de dialogue actuels, le traitement d'un grand nombre de figures de style est négligé, comme nous allons le voir maintenant. La section suivante présente un certain nombre de modèles définissant les mécanismes utiles à la résolution des références, et en particulier des références multimodales.

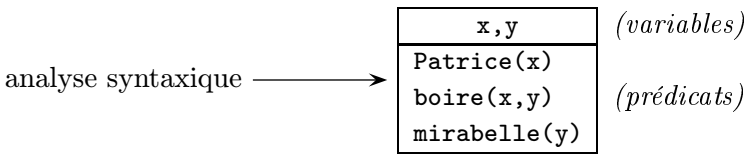
## 2.2 Méthodes de résolution

### 2.2.1 DRT

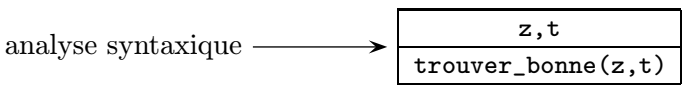
La DRT *(Discourse Representation Theory)* de [Kamp 88] propose de traiter de manière logique le discours. Elle consiste à construire une structure, la DRS *(Discourse Representation Structure)*, qui va contenir les référents sous la forme de variables et les actions sous la forme de prédicats. Cette structure permet de définir des contextes d'accessibilité, en particulier pour la résolution des *anaphores*, qui se fait par *unification des variables*.

Construction de la DRS sur l'exemple : *"Patrice boit de la mirabelle. Il la trouve bonne."*

**Premier énoncé :** *"Patrice boit de la mirabelle."*




**Deuxième énoncé :** *"Il la trouve bonne."*



unification des variables selon des critères syntaxiques : ici, le *genre* permet de trouver la solution →

| x,y,z,t |
|---|
| Patrice(x) |
| boire(x,y) |
| mirabelle(y) |
| z=x |
| t=y |
| trouver_bonne(z,t) |

La DRT comprend également des règles de traitement pour la négation, la disjonction, l'implication, ainsi que pour la quantification et certains aspects temporels du discours. Des domaines de validité sont parfois utilisés pour limiter la portée des reprises anaphoriques. Ainsi, un référent utilisé dans une proposition négative ne peut pas être repris.

La DRT ne tient compte que du discours et ne permet donc pas de traiter les références multimodales.

### 2.2.2 Axiologie

[Gaiffe 92] définit une axiologie comme une relation binaire entre des propriétés opposées se déroulant nécessairement à l'intérieur d'un domaine de référence. Lorsque le discours énonce une catégorie d'objet N, une opposition se crée entre les objets de l'environnement qui sont de la catégorie N, et ceux qui sont d'une autre catégorie. Lorsqu'une propriété P est en jeu, il se crée de même une opposition entre les objets de la catégorie N qui ont la propriété P, et les autres objets de la catégorie N. Cette opposition pourra être à l'origine d'une reprise ultérieure (par exemple : *"les triangles rouges"* puis *"les autres"*). C'est l'un des principaux intérêts du modèle axiologique qui permet, lorsque les schémas axiologiques décrits dans le schéma ci-dessous sont sauvegardés dans l'historique du discours, de traiter ces reprises. Selon le type de déterminant, un domaine de référence (c'est-à-dire un cadre dans lequel l'opposition va s'appliquer) est créé et l'opposition est construite de la manière suivante :

| Défini | Démonstratif | Indéfini |
|---|---|---|
| dans un domaine de référence Q contenant des N et des non-N<br><br>*"le N"* ou *"les N"* :<br><br>Q<br>N → non N | dans un domaine de référence contenant des N pouvant avoir ou ne pas avoir la propriété P<br><br>*"ce N"* ou *"ces N"* :<br><br>N<br>ce N — autres N<br>P → non P | dans un domaine de référence Q contenant des N et des non-N, les N pouvant avoir ou ne pas avoir la propriété P<br><br>*"un N"* ou *"des N"* :<br><br>Q<br>N → non N<br>un N — d'autres N<br>P → non P |

Dans un énoncé tel que *"ce N"*, aucune propriété P n'est explicite. C'est en effet l'emploi

du démonstratif qui sous-entend une propriété P. D'une manière générique, P veut dire saillant. Ainsi, *"ce N"* sera saillant soit parce qu'il est singularisé dans le discours, soit parce qu'il est désigné par un geste. Face à un démonstratif, le modèle axiologique permet en effet de considérer le geste, mais il ne fournit pas de règles précises permettant de choisir entre une anaphore et une référence multimodale. A priori, seule la présence d'un geste permet de supposer que la propriété saillante est avant tout celle d'être désigné par le geste.

### 2.2.3 Représentations mentales

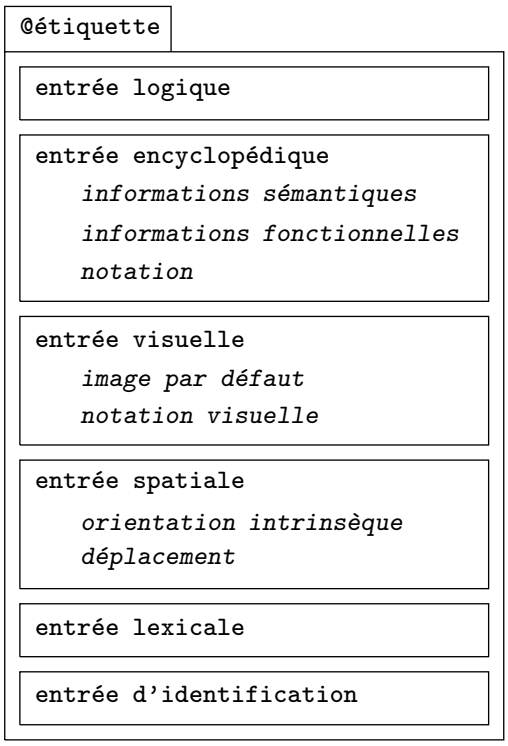
Les Représentations Mentales ou RM est un modèle de représentation des référents proposée par [Reboul 98]. Une RM est attachée à chaque objet de l'application et contient, dans la mesure du possible, l'ensemble des informations attachées à cet objet et susceptibles d'être activées lors d'une référence à cet objet.

Une RM se présente comme une suite d'informations regroupées sous divers champs, avec une **étiquette** qui permet de la reconnaître ou de la désigner. Les champs et sous-champs sont les suivants :

- Une **entrée logique** qui rassemble les différentes relations logiques que la RM entretient avec d'autres RM, en particulier lorsque plusieurs RM forment un groupement comme nous allons le voir ci-dessous.
- Une **entrée encyclopédique** qui rassemble toutes les informations dont dispose l'individu sur le référent, hors celles qui entrent dans les autres champs de la RM. Ainsi, les *informations sémantiques* regroupent les informations conceptuelles (au niveau sémantique) que la RM hérite de la catégorie à laquelle le référent correspond. De même, les *informations fonctionnelles* indiquent à quoi sert l'objet. Un troisième type d'informations, appelé *notation*, regroupe les informations spécifiques à l'objet correspondant à la RM.
- Une **entrée visuelle** qui rassemble les informations visuelles : l'*image par défaut* qui correspond aux informations héritées des connaissances générales sur le monde, et la *notation visuelle* qui correspond aux informations tirées de la perception effective de l'objet.
- Une **entrée spatiale** qui rassemble les informations spatiales sur l'objet : l'*orientation intrinsèque* de l'objet s'il en a une, la position de l'objet par rapport aux autres objets, la trace de ses *déplacements*.
- Une **entrée lexicale** qui rassemble les informations linguistiques élémentaires : expressions référentielles effectivement utilisées ou qui pourraient être utilisées pour désigner l'objet.
- Une **entrée d'identification** qui indique l'identification de l'objet pour l'application dans le dialogue homme-machine.

Afin de gérer les ensembles d'objets que le discours fait apparaître, ainsi que les groupements perceptifs que le geste fait apparaître, de nombreuses opérations sur les RM sont définies. Citons en particulier :

- la **création** d'une RM lorsqu'un nouvel objet est mentionné dans le discours ou perçu visuellement ;
- la **fusion** de deux RM lorsqu'elles correspondent au même objet ;
- le **groupement** de deux RM lorsque le discours contient par exemple une énumération ou une coordination, ou lorsqu'un groupe perceptif apparaît ;




– l'**extraction** d'une RM lorsque par exemple le discours ou le geste reprend un élément d'un ensemble.

Enfin, divers traitements sont proposés selon le type d'expression référentielle contenue dans le discours. La spécification de ces traitements a nécessité, dans la mesure du possible, l'identification de tous les types d'expressions possibles et des mécanismes associés. Ainsi, non seulement les pronoms, les groupes nominaux définis et indéfinis impliquent leurs propres règles, mais également les adjectifs ordinaux, les mots "suivant", "dernier" et "autre".

Le modèle des RM est en cours de développement et c'est pourquoi nous n'en parlerons pas plus. Il semble cependant que, étant le seul modèle à tenir réellement compte des informations extra-linguistiques comme celles données par la perception visuelle, il soit bien adapté à la résolution de la référence multimodale.

### 2.2.4 Autres méthodes de résolution des références multimodales

Pour faire face à des impératifs liés à l'implantation des systèmes de dialogue, plusieurs méthodes de résolution des références multimodales ont été proposées. Rapidement implantables et souvent efficaces pour la majorité des énoncés multimodaux, elles se révèlent néanmoins très insuffisantes face à la diversité du langage naturel et du geste spontané.

Ainsi, l'énoncé oral et l'énoncé gestuel sont souvent considérés comme deux filtres dont la combinaison permet de trouver une solution à la référence multimodale :

- Une expression référentielle est un **filtre linguistique** dans le sens que le substantif et les éventuels adjectifs donnent des caractéristiques permettant de délimiter un ensemble d'objets. Seuls les objets appartenant à cet ensemble seront considérés lors de la résolution de la référence multimodale. Le substantif correspond généralement à la catégorie, et les adjectifs donnent des informations sur certaines caractéristiques perceptives des objets ou

sur leur nombre. Le filtre n'est parfois construit que sur la catégorie seule, ce qui est pour le moins simpliste et réducteur.

- Un geste de désignation est un **filtre perceptif** dans le sens qu'il délimite un sous-espace perceptif. Seuls les objets appartenant à ce sous-espace seront considérés.

La combinaison de ces filtres peut se faire de plusieurs façons :

(1) Les deux filtres sont appliqués aux objets de l'application chacun de leur côté, les deux ensembles d'objets obtenus étant ensuite intersectés.
(2) Le filtre *linguistique* est appliqué aux objets de l'application, puis le filtre *perceptif* est appliqué sur l'ensemble d'objets obtenu.
(3) Le filtre *perceptif* est appliqué aux objets de l'application, puis le filtre *linguistique* est appliqué sur l'ensemble d'objets obtenu.

S'il est clair que la première méthode n'est pas optimale, le choix entre la deuxième et la troisième s'avère plus délicat. Le filtre linguistique nécessite un traitement plus complexe que le filtre perceptif, qui n'est en fait qu'une série de comparaisons de coordonnées. La troisième méthode s'avèrera donc plus rapide que la deuxième, voire incontournable si l'application met en jeu un très grand nombre d'objets.

En considérant l'ambiguïté de portée du geste, il apparaît cependant que le filtre perceptif, contrairement au filtre linguistique, peut varier. Ainsi, si aucun objet ou si trop d'objets sont trouvés, la variation du filtre perceptif permet d'aboutir de proche en proche à un résultat adéquat. Or cette variation n'est intéressante qu'une fois le filtre linguistique appliqué, c'est-à-dire seulement dans la deuxième méthode. On en déduit que, dans les applications pour lesquelles le geste risque d'être imprécis, la deuxième méthode s'avère nécessaire. C'est le cas en particulier dans les applications en *3D* et dans certaines applications en *2D* pour lesquelles la superposition d'objets est fréquente : lorsqu'un objet A cache un objet B, l'intention de désigner B se traduit par un geste sur A. Si l'énoncé oral contient la description de B et si le filtre perceptif ne s'étend pas jusqu'à inclure B, la référence multimodale ne sera pas résolue.

Cette méthode consistant à étendre un ensemble d'objets candidats jusqu'à obtenir un résultat possible est une méthode classique de résolution des références : on part de l'ensemble le plus réducteur possible, par exemple le *focus* (c'est-à-dire le dernier objet manipulé), puis on étend l'ensemble jusqu'à ce qu'il contienne tous les objets de l'application. Le traitement s'arrête lorsque l'ensemble contient un résultat possible à la référence, c'est-à-dire des candidats ayant les caractéristiques requises et étant en nombre requis. Les étapes sont par exemple :

1. le focus,
2. les objets présents dans l'historique du dialogue,
3. les objets saillants visibles à l'instant de la référence (c'est-à-dire les objets apparaissant à l'écran et ayant des caractéristiques perceptives qui les rendent facilement repérables),
4. tous les objets visibles à l'instant de la référence,
5. tous les objets de l'application.

Dans cette liste ordonnée, les contextes linguistiques se trouvent en amont des contextes perceptifs. La priorité de la mémoire du discours sur la perception visuelle est un argument que l'on trouve par exemple dans [Moulton 94]. Encore une fois, le principe même de cette méthode consistant à essayer d'obtenir un résultat sans vraiment comprendre le mécanisme de référence s'avère totalement insuffisant face aux subtilités de la langue naturelle et du geste spontané.

# Deuxième partie

# Analyse des références multimodales dans un corpus et modélisation théorique

# Introduction : Démarche adoptée

### Objet de l'analyse

Comme nous l'avons vu dans l'introduction générale, la première étape de notre travail consiste en l'analyse de la référence multimodale dans un corpus. Ce corpus multimodal a été constitué par Frédéric Wolff [Wolff 99] lors d'une expérience mettant des sujets face à une application de type *commande de processus*. La seule tâche possible de l'application était le rangement d'objets dans des lieux précis ; les énoncés multimodaux attendus devaient donc comporter des désignations d'objets et des désignations de lieux, désignations soit uniquement orales, soit multimodales.

L'objectif de Frédéric Wolff était de recueillir la plus grande variété possible de gestes de désignation, afin d'étudier leurs formes et les utilisations de celles-ci selon le contexte perceptif visuel, c'est-à-dire selon la disposition des objets dans la scène et les uns par rapport aux autres. Notre objectif est l'étude des références multimodales. Nous nous intéresserons ainsi à la synchronisation entre parole et geste, à la fréquence des apparitions du geste en fonction des expressions référentielles, ainsi qu'aux rapports entre les différents types de gestes et les différents types d'expressions référentielles. Bien que notre objectif diffère totalement de celui à l'origine de l'expérience, le corpus présente un grand intérêt car les sujets ont largement utilisé la parole, permettant ainsi le recueil d'un grand nombre d'expressions référentielles représentatives de la langue naturelle.

### Démarche d'analyse du corpus

Le but de l'analyse est d'identifier les problèmes apparaissant lors de la compréhension des références multimodales. Une première étape consistera à repérer toutes les références multimodales, à identifier et à classer :

- les types de synchronisations entre les gestes et les expressions référentielles ;
- les types de gestes et les informations nécessaires pour leur compréhension ;
- les types d'expressions référentielles, les types de données qu'elles contiennent, et les informations nécessaires pour leur compréhension ;
- les types de relations entre les gestes et les expressions référentielles.

Nous calculerons la fréquence d'apparition dans le corpus des phénomènes identifiés, afin d'évaluer l'importance de chacun et de focaliser l'étude sur les problèmes les plus fréquents.

Ensuite, l'analyse consistera à identifier les différents types d'ambiguïtés, c'est-à-dire les différentes manières selon lesquelles les imprécisions du geste ou de l'énoncé oral peuvent se combiner pour aboutir à un énoncé ayant plusieurs interprétations possibles dans le contexte de sa production. Nous nous intéresserons aux informations qui permettent à un interlocuteur humain de comprendre ces énoncés et nous étudierons les possibilités de traitement de ces informations.

# Chapitre 3

# Analyse du corpus multimodal MagnétOz

## 3.1 Présentation de l'expérience MagnétOz et analyse globale

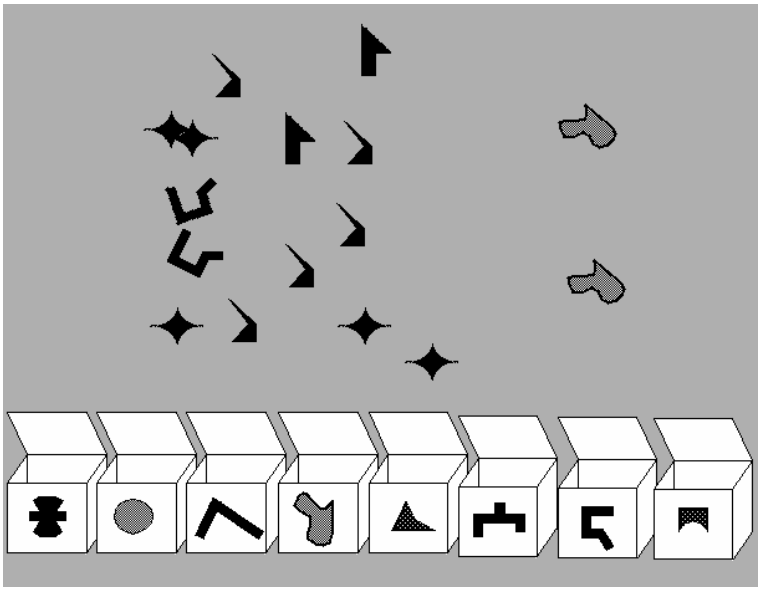
### 3.1.1 Objectifs et mise en œuvre

Cette expérience a été mise en place par Nadia Bellalem et Frédéric Wolff, dans le cadre de la thèse de ce dernier [Wolff 99]. L'objectif était de recueillir le plus grand nombre possible de gestes effectués en situation de communication homme-machine à l'aide d'un écran tactile (le dispositif de vis-à-vis le plus naturel). Or le recueil de données correspondant à une situation de communication homme-machine pose un problème : ces données sont généralement recueillies afin de préparer l'implantation d'un système qui n'existe pas encore. La communication homme-machine doit donc être simulée : les sujets de l'expérience croient qu'ils communiquent avec une machine mais c'est en fait un homme (le magicien ou compère) qui dirige le pseudo-système. Ce type d'expérience est appelée Magicien d'Oz en référence au roman de Baum. Lors de l'expérience MagnétOz, Frédéric Wolff a tenu le rôle du magicien.

Afin de recueillir des gestes, le sujet doit être incité à utiliser fortement le geste. L'application suivante a ainsi été définie : il s'agit de ranger des objets dans des lieux qui leur sont attribués, les objets et les lieux étant susceptibles d'être désignés par la voix, par le geste, ou par les deux. Dans un premier type de scènes (en *2D*), les objets à ranger ne sont pas nominalisables, c'est-à-dire que leurs formes ont été choisies de manière à rendre très difficile l'usage de la parole seule. Au contraire, dans un second type de scènes (en *2D* également), les objets sont des jouets (peluches, ballons et modèles réduits de voitures) et les lieux sont des meubles de rangement (un coffre pour les ballons et des étagères pour les autres jouets). Ceci permettra de comparer la production du geste dans les deux cas.

Comme le montrent les copies d'écran suivantes, les objets sont dispersés dans la scène de manière apparemment aléatoire. Ils sont en fait groupés ou isolés selon des critères liés à la perception visuelle. Des groupes perceptifs dans le sens de la proximité sont ainsi élaborés, de même que des objets isolés. Afin de perturber la production des gestes et d'inciter le sujet à varier ses gestes, des distracteurs (objets qui ne sont pas à ranger) viennent de temps en temps perturber les groupes perceptifs.

L'expérience a été réalisée sur 7 sujets non-informaticiens, à raison d'une séance d'environ 30 minutes par sujet. Le sujet était installé à un poste comprenant un micro et un écran tactile qu'il devait utiliser avec un stylet et qui n'affichait pas la trajectoire des gestes. Ce poste trans-



mettait (par le réseau) les sons et les images au magicien installé dans une autre pièce. Celui-ci commandait le pseudo-système à l'aide de quelques messages pré-enregistrés et en déplaçant en temps réel les objets de l'application. Au cours de chaque séance, environ une centaine d'énoncés étaient produits, à raison de 2 à 5 par scène, les scènes étant réparties de la manière suivante : environ 20 avec des formes et 5 ou 6 avec des jouets.

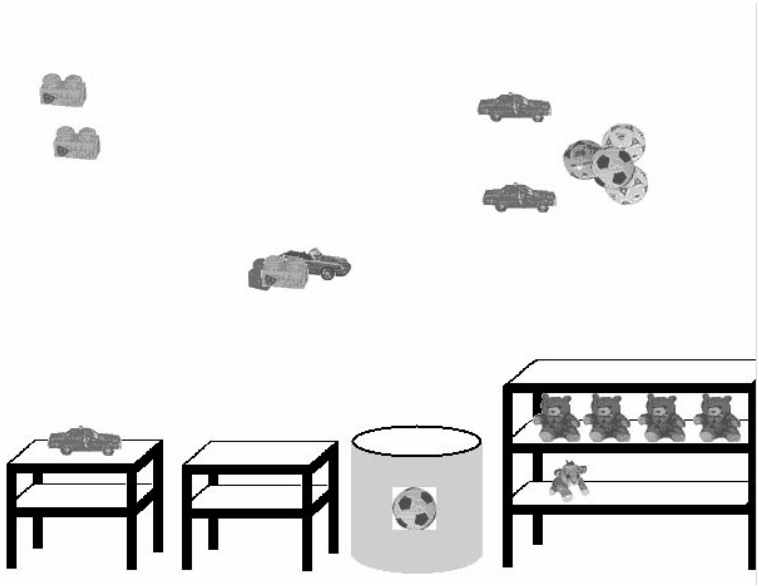
Les sons et images étaient également enregistrés dans des fichiers qui constituent le corpus. Un fichier XML contenant des pointeurs sur ces fichiers a ensuite été généré. Plus précisément, Frédéric Wolff a développé un logiciel permettant d'une part d'exécuter l'expérience (logiciel serveur pour le magicien et logiciel client pour le sujet), d'autre part de l'enregistrer puis de la relire. Les images ne sont pas enregistrées telles quelles mais sous la forme de trajectoires gestuelles et de configurations des objets à chaque instant. Le logiciel MagnétOz reconstruit les images à partir de ces données et offre la possibilité de naviguer d'une scène à l'autre.

### 3.1.2 Analyse globale du corpus

L'étude des gestes étant l'objectif de cette expérience, le corpus n'a tout d'abord été étudié que de ce point de vue. Il présente pourtant un grand intérêt du point de vue de la parole (grande diversité compte tenu du caractère restrictif de la tâche) et donc de la multimodalité. Notre premier travail a ainsi été la relecture de ce corpus d'un point de vue linguistique (étude des énoncés oraux, de leurs structures, de leurs expressions référentielles) puis d'un point de vue multimodal (étude de l'articulation entre parole et geste lors des références).

Une première analyse du corpus permet de faire les observations suivantes :

- **Déroulement du dialogue :**
  Au début de chaque séance, le magicien décrivait la tâche de rangement à accomplir. Lors de l'affichage de chaque nouvelle scène, le magicien indiquait lorsqu'il était prêt et le sujet énonçait alors sa commande. Lorsque le magicien ne comprenait pas un énoncé (geste très mal fait, voix inaudible, ou surtout présence d'une ambiguïté), il effectuait soit une demande de reformulation de l'énoncé complet, soit une demande de précision sur les objets ou sur le lieu. Cette logique a toujours été suivie, du début jusqu'à la fin des séances (par



exemple, le sujet a toujours attendu que le magicien lui dise qu'il est prêt avant d'énoncer quelque chose).

Les énoncés sont quasiment tous indépendants : il n'y a pratiquement jamais d'anaphore ni de reprise entre deux tours de parole. De même, du fait de la tâche de rangement, une action ne s'est jamais effectuée sur plusieurs tours de parole. En règle générale, le magicien a toujours compris l'intention du sujet et a effectué l'action correspondante, même si elle était décrite maladroitement. On ne trouve qu'un exemple dans le corpus où l'intention du sujet n'est pas comprise par le magicien : le sujet a beau répéter 4 fois son énoncé ou un énoncé proche, l'incompréhension est totale et n'aboutit pas à l'exécution de la tâche (qui était d'ailleurs contraire au but défini). Le sujet est même allé jusqu'à faire des essais sur des distracteurs, en parallèle de la tâche en cours.

- **Types d'énoncés :**

  Tous les énoncés sont des ordres de rangements. Ils prennent pourtant plusieurs formes, souvent différentes de l'impératif ou de l'infinitif. On trouve par exemple les tournures : *"Je mets ces objets dans cette boîte"*, *"Tu dois ranger cet objet ici"*, *"Ces formes se rangent ici"*, *"Ces objets iront dans cette boîte"*, *"Cette figure doit être rangée ici"*, *"Ces objets ici"*, *"Peux-tu dissocier ces objets ?"*, etc.

  La multimodalité est utilisée de manière très variable selon les sujets. Parmi les sept sujets, deux vont systématiquement faire des pointages précis et utiliser des expressions référentielles très simples. Un autre, par contre, va systématiquement essayer de ne pas faire de geste et de décrire précisément les objets. Ce sujet est d'ailleurs le plus difficile à comprendre. Les autres sujets montrent une grande diversité et passent d'une série d'énoncés purement oraux à des énoncés utilisant efficacement la multimodalité. Ce sont eux qui sont les plus intéressants car leurs gestes et leurs expressions référentielles varient très souvent (on retrouve tous les types d'ambiguïtés possibles).

  D'autre part, on observe dans certains énoncés :

  – un geste qui n'est pas en synchronisation avec l'expression référentielle à laquelle il se rapporte (parfois quelques secondes avant, parfois quelques secondes après),

  - un voire deux gestes involontaires,
  - une répétition de l'expression référentielle, qui s'accompagne parfois d'une répétition d'un geste (mais pas toujours).

- **Types de désignations de lieu :**
  Le lieu est presque toujours la destination du rangement, bien que dans quelques cas, c'est un coin vide de la scène qui est désigné. Le sujet va alors y regrouper des objets ou, au contraire, écarter un distracteur du groupe d'objet qu'il va ensuite pouvoir ranger à l'aide d'un accès pluriel.
  Les lieux sont souvent décrits par une expression référentielle contenant suffisamment d'informations, c'est-à-dire par exemple le rang de la boîte avec le sens du décompte. Un geste de désignation est parfois présent (il est alors redondant).
  Les gestes de désignation de lieu sont le plus souvent des pointages, parfois des gribouillages ou des pointages très étendus (dénotant dans certains cas le trajet que les objets doivent parcourir pour être rangés).

- **Types de désignations d'objets :**
  Du fait de la tâche, les sujets cherchent à ranger tous les objets d'un même type dans leur boîte en une seule commande. Si les objets forment un groupe perceptif au sens de la proximité, un geste global est fréquent, de même qu'une expression référentielle au pluriel. Ce n'est cependant pas toujours le cas : chaque objet peut être désigné séparément et l'énoncé contient alors une suite de références multimodales. De même, il arrive qu'un geste global désigne des objets qui ne forment pas de groupe perceptif (du fait de la présence de distracteurs).
  Lorsqu'un geste est effectué, les expressions référentielles les plus fréquentes sont des groupes nominaux démonstratifs. On trouve cependant avec le geste un grand nombre de groupes nominaux définis. Les gestes les plus fréquents pour un accès singulier sont des pointages, et les gestes les plus fréquents pour un accès pluriel sont des entourages. On trouve cependant quelques pointages ou gribouillages sur des groupes perceptifs, ainsi que quelques entourages sur un objet isolé.

## 3.2 Analyse des références aux objets dans le corpus MagnétOz

### 3.2.1 Expressions référentielles

Une expression référentielle peut contenir beaucoup plus qu'un déterminant et la catégorie du ou des objets référés. En particulier lorsqu'il ne fait pas de geste, le sujet peut décrire certaines caractéristiques des objets de manière très précise, ou bien donner des indications permettant d'identifier ces objets. La présence de ces informations est très aléatoire, et, chez un même sujet, on trouve aussi bien des expressions sans aucune information que des expressions combinant plusieurs types d'informations de manière parfois redondante. Ainsi, on trouve dans le corpus :

- Des expressions référentielles ne contenant **aucune information** : *"ça"* n'indique absolument rien, ni le nombre des référents ni le genre de leur catégorie.

- Des indications sur le **mode de référence** et sur le nombre des référents : on trouve en effet des déterminants définis, démonstratifs, des pronoms personnels représentants, des pronoms démonstratifs, quelques déterminants numéraux cardinaux, quelques (rares) déterminants

indéfinis, quelques (très rares) déterminants quantifiants indéfinis. Un déterminant ou pronom au singulier indique un accès individuel. Un déterminant ou pronom au pluriel indique soit un accès pluriel, soit un accès multiple. La distinction singulier/pluriel n'est pas toujours indiquée par le déterminant : *"celle(s)-là"* se prononce de la même façon au singulier et au pluriel. Un marqueur déictique est souvent associé au groupe nominal démonstratif et souvent également au groupe nominal défini, ce qui constitue dans ce dernier cas une forme grammaticale incorrecte *("l'objet-là")*. D'autre part, une indication de présence d'un geste de désignation peut être explicite dans l'énoncé : *"les objets pointés"*.

- Le **nombre des référents**. Il peut être indiqué explicitement par un adjectif numéral, ce qui est souvent le cas : *"ces deux objets"*. Il peut aussi se déduire d'autres informations indiquées dans l'expression : *"ces objets formant un triangle"* (cet exemple est un cas unique dans tout le corpus).

- La **catégorie des référents**. Dans les scènes où les objets à ranger sont des formes difficiles à nominaliser, on trouve pas moins de 11 substantifs (éventuellement des mots composés) : *"objet"*, *"forme"*, *"forme géométrique"*, *"figure"*, *"pièce"*, *"rond"*, *"cercle"*, *"triangle"*, *"pointe"*, *"flèche"*, *"bout de flèche"*. Dans les scènes où les objets à ranger sont des jouets, on trouve : *"nounours"*, *"ours"*, *"ourson"*, *"peluche"*, *"animal"*, *"balle"*, *"ballon"*, *"ballon de foot"*, *"voiture"*, *"jouet"*. Les peluches se distinguaient en deux catégories : les ours en peluche et les lions en peluche. On remarque que tous les sujets n'ont référé un lion en peluche que par un hyperonyme comme *"peluche"*. On en déduit que la représentation graphique du lion en peluche ne permettait probablement pas d'identifier facilement l'animal. D'autre part, certains substantifs portent en eux une certaine ambiguïté. Ainsi, *"étagère"* peut aussi bien référer au meuble complet qu'à une seule planche du meuble, parfois chez le même sujet. *"objet"*, dans certains cas relativement rares, peut référer aussi bien à un seul objet qu'à un groupe d'objets constituant une entité unique (d'où le singulier). L'expression *"cet objet"* est même, dans un cas unique du corpus, utilisé deux fois dans le même énoncé, la première fois pour désigner un objet unique, la deuxième fois pour désigner un groupe d'objets.

- Des **informations liées à l'évolution de la tâche** : *"suivant"*, *"restant"*, *"autre"*, *"premier"*, *"dernier"*. Ces informations ne sont données que rarement. Elles peuvent correspondre soit au contexte discursif, soit au contexte perceptif.

- Des **informations liées au contexte perceptif** :
  - Sur l'**homogénéité des caractéristiques visuelles des référents** : *"ces objets de même couleur"*.
  - Sur une **caractéristique visuelle des référents** : *"gris"*, *"gris clair"*, *"clair"*, *"en pointillés"*, *"à petits pois"*, *"qui comporte un angle droit"*, *"en forme de L"*, *"en forme de hache"*.
  - Sur une **caractéristique visuelle des référents en comparaison aux autres objets** : *"les trois formes les plus claires"*.
  - Sur la **disposition des référents entre eux** : *"cette ligne d'objets"* (simplifié ensuite en *"cette ligne"*), *"ces objets formant un triangle"*, *"les trois objets groupés"*, *"ce groupement d'objets"* (information précisant la cohésion des référents).
  - Sur la **disposition des référents par rapport aux autres objets** : *"l'objet le plus à droite"*.
  - Sur la **disposition des référents dans la scène** : *"les deux objets au centre"*.

### 3.2.2 Ambiguïté des références multimodales

Dans cette section et dans la suite, nous considérons qu'une référence multimodale est ambiguë lorsque les théories classiques (intersection du filtre linguistique et du filtre perceptif) donnent soit trop d'objets candidats à la référence multimodale, soit aucun objet.

Afin de décrire les ambiguïtés des références multimodales, nous identifions la manière dont une ambiguïté sur une modalité est affectée par l'autre modalité. Lorsque c'est nécessaire, la signification du terme *ambigu* est explicitée :

1. **Références multimodales non ambiguës :**
   - ***L'énoncé oral et le geste non ambigus et en adéquation :*** l'énoncé oral décrit des objets qui peuvent être identifiés (en tenant compte de l'énoncé oral seul) sans aucun problème ; le geste désigne les mêmes objets sans aucune ambiguïté (ni de forme ni de portée) ; et l'énoncé multimodal n'est donc pas ambigu. C'est le cas par exemple de : "la voiture" + geste sur la seule voiture de la scène. Les informations données par les modalités sont redondantes.
   - ***Les ambiguïtés de l'énoncé oral résolues par le geste :*** c'est le cas en particulier de l'interférence entre contexte discursif et contexte perceptif. En effet, une expression démonstrative peut être aussi bien une anaphore qu'une désignation. La présence d'un geste non ambigu permet de vérifier et valider la possibilité de la désignation. Un autre exemple correspond à l'ambiguïté linguistique du pluriel indéterminé : "ces objets" + geste sur trois objets.
   - ***Les ambiguïtés du geste résolues par l'énoncé oral :*** lorsque le geste est ambigu et ne permet pas de déterminer le nombre d'objets désignés (ambiguïté de portée), l'énoncé oral peut donner la solution s'il contient le nombre de référents. C'est le cas de l'exemple : "les deux objets" + geste imprécis sur deux ou trois objets. De même, la catégorie indiquée dans l'énoncé oral peut permettre de résoudre un geste imprécis sur un amas d'objets, comme dans l'exemple : "la voiture" + geste sur un amas d'objets contenant une seule voiture.
   - ***Les ambiguïtés de l'énoncé oral et du geste résolues mutuellement :*** en considérant les exemples donnés précédemment, une expression démonstrative comportant un nombre d'objet et un geste comportant une ambiguïté de portée peuvent s'associer et aboutir à une référence multimodale non ambiguë. C'est le cas de l'exemple : "ces deux objets" + geste imprécis sur deux ou trois objets. Un autre exemple plus intéressant fait intervenir d'un côté le nombre d'objets, de l'autre côté la catégorie des objets : "ces triangles" + geste ambigu de trois ou quatre objets, dont trois triangles sûrs.

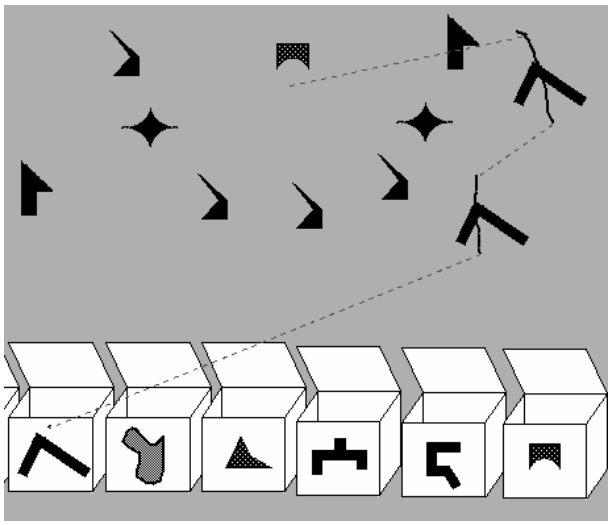
2. **Références multimodales ambiguës :**
   - ***L'énoncé oral et le geste non ambigus et en inadéquation :*** ce cas est extrêmement rare dans le corpus et correspond à chaque fois à une erreur du sujet.
   - ***Les ambiguïtés de l'énoncé oral non résolues par le geste :*** ce cas est également anecdotique et fait intervenir des mécanismes que nous étudierons dans le chapitre suivant. Il s'agit par exemple de : "les objets de cette forme" + geste non ambigu sur un objet, l'ambiguïté linguistique sur le pluriel indéterminé n'étant pas résolue par le geste puisque celui-ci désigne seulement un des objets référés.

- ***Les ambiguïtés du geste non résolues par l'énoncé oral :*** encore une fois, ce cas est anecdotique : "les trois objets de la même forme" + geste sur six objets, trois ayant la même forme, et les trois autres ayant également la même forme (mais étant en nombre supérieur à trois si on considère l'espace perceptif complet).
- ***Les ambiguïtés de l'énoncé oral et du geste non résolues mutuellement :*** il s'agit ici de véritables ambiguïtés multimodales. Contrairement aux trois types de références précédents, l'ambiguïté multimodale est présente à plusieurs endroits dans le corpus. C'est le cas de l'exemple : "ces objets" + geste sur deux ou trois objets. Un autre exemple plus subtil fait intervenir une ambiguïté sur le singulier ou le pluriel : "celle(s)-ci" + geste sur une ou deux formes.

### 3.2.3 Références multimodales combinées

Lorsque plusieurs références multimodales se suivent dans un même énoncé, on peut assister à un phénomène que nous avons appelé *référence multimodale combinée* et qui consiste en l'association d'une désignation dans une modalité à plusieurs désignations dans l'autre modalité.

Le schéma suivant correspond à l'exemple du corpus : "ces deux objets" + deux gestes, un pour chaque objet :

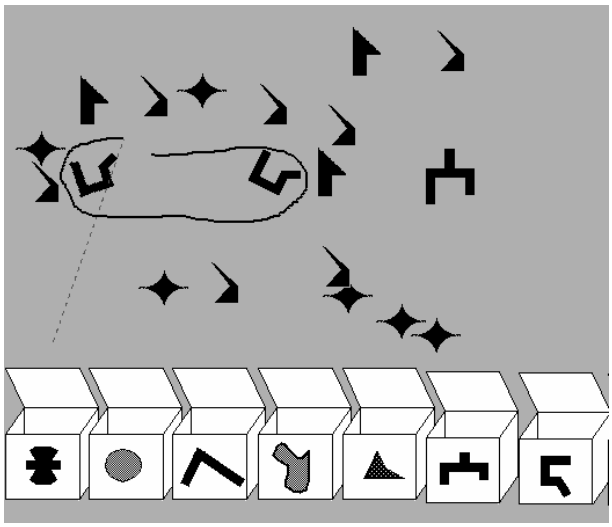


Au contraire, le schéma de la page suivante correspond à l'exemple : "cet objet et celui-ci" + un geste pour les deux objets.

Une première classification des formes que peut prendre ce phénomène se base sur le nombre d'expressions référentielles et le nombre de gestes (les quatre possibilités sont présentes dans le corpus) :

- une expression référentielle est associée à un geste,
- une expression référentielle est associée à plusieurs gestes,
- plusieurs expressions référentielles sont associées à un geste,
- plusieurs expressions référentielles sont associées à plusieurs gestes.

En considérant que l'interprétation de "ces deux objets" est comparable à celle de "cet objet et celui-ci", une deuxième classification se base sur le nombre d'objet référés par l'énoncé oral



et le nombre de gestes. Ainsi, dans l'énoncé "cet objet, celui-ci et celui-ci" + deux gestes (un geste de pointage sur un objet puis un geste d'entourage de deux objets), l'énoncé oral indique le nombre d'objets référés (trois), et c'est à partir de ce nombre que les gestes vont être interprétés : on cherche trois objets, et, comme on est en présence de seulement deux gestes, on va chercher à interpréter l'un des gestes comme un accès pluriel à deux objets. La structure de l'énoncé oral en trois expressions référentielles n'est qu'un critère secondaire, qui peut être intéressant par exemple si les gestes sont bien synchronisés avec les expressions auxquelles ils se rapportent.

Les possibilités d'associations de $n$ référents (selon l'énoncé oral) avec $p$ gestes sont les suivantes (elles sont toutes présentes dans le corpus) :

| *référents linguistiques* | *nombre de gestes* | *types de gestes* |
|---|---|---|
| n | 1 | pluriel |
| n | 1<p<n | indifférent |
| n | n | individuel |
| indéterminé | 1 | pluriel |
| indéterminé | 1<p<n | indifférent |
| indéterminé | n | individuel |

Un système qui se trouve face à de tels énoncés doit retrouver les associations. Cette tâche est d'autant plus complexe que :

- la synchronisation entre gestes et expressions référentielles n'est pas toujours très bonne (il peut se passer plusieurs secondes après la fin de l'énoncé avant que le geste ne commence),
- la désignation d'objets et la désignation de lieu peuvent interférer,
- une expression référentielle peut être répétée, précisée ou corrigée par le sujet,
- un geste peut être répété,
- un geste peut être involontaire.

Le chapitre suivant propose d'une part un algorithme dont le but est de retrouver les associations, d'autre part une synthèse théorique sur l'interprétation des références multimodales.

# Chapitre 4

# Interprétation d'énoncés multimodaux

## 4.1 Traitement des références multimodales combinées

### 4.1.1 Identification du problème

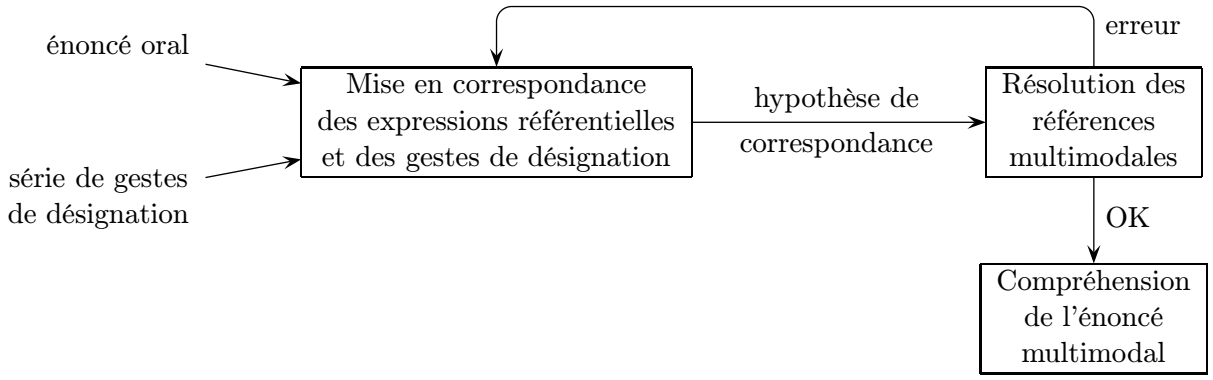
Ce qui ne pose aucun problème de compréhension pour un interlocuteur humain peut au contraire s'avérer un problème extrêmement complexe pour un système informatique. C'est le cas de la distinction entre geste désignant un objet et geste désignant un lieu. Nous avons vu que dans le corpus MagnétOz, un énoncé multimodal contient généralement une ou plusieurs désignations gestuelles d'objets et parfois une désignation gestuelle de lieu. Lors de la référence au lieu et dans le cas d'une expression référentielle définie non ambiguë, aucun critère ne permet de prévoir si un geste va être effectué ou non.

Prenons un exemple très simple d'énoncé multimodal (14). L'ambiguïté entre d'un côté deux gestes pour les objets, de l'autre côté un geste pour les deux objets suivi d'un geste pour le lieu, est totale. Ni la synchronisation ni le contexte applicatif ne permettent de faire un choix sûr. Cette ambiguïté ne se pose cependant jamais pour le magicien, qui va accepter et traiter l'énoncé sans se poser de questions. En effet, l'interlocuteur humain est capable de déterminer immédiatement si un geste désigne un objet ou un lieu, selon des critères perceptifs et logiques très complexes. Par exemple, le deuxième geste peut être interprété comme une désignation de lieu parce que les objets les plus proches de la trajectoire ne forment pas de groupe perceptif avec le ou les objets désignés par le premier geste. C'est une raison parmi beaucoup d'autres possibles. Elle est saillante pour l'interlocuteur humain mais un système aura beaucoup de mal à faire la part des choses entre toutes les possibilités.

(14) *"Range ces deux objets dans la troisième boîte"* + deux gestes

Ainsi, le système n'aura aucun moyen de distinguer a priori un geste de désignation d'objet d'un geste de désignation de lieu. Comme de plus la synchronisation n'est pas souvent parfaite et qu'il est nécessaire de tenir compte des références multimodales combinées, se pose le problème consistant à établir les correspondances entre gestes de désignation et expressions référentielles. Ce problème est distinct de celui consistant à résoudre les références multimodales. Or, dans les approches classiques, soit le calcul de référents n'est effectué que sur les références multimodales simples, soit les deux problèmes sont confondus.

Nous en déduisons une première proposition d'architecture pour un système de traitement des énoncés multimodaux (schéma page suivante). Dans le cadre de ce stage de DEA, notre objectif est de décrire le problème de mise en correspondance que nous avons identifié, c'est-à-dire de spécifier les informations en entrée et leur traitement.




### 4.1.2 Informations nécessaires en entrée

Nous nous plaçons dans le cadre d'un seul énoncé multimodal, donc tout ce qui suit est valable à la fois pour un système de dialogue *en temps réel* et pour l'étude a posteriori du corpus MagnétOz. Les informations nécessaires en entrée du module de traitement peuvent être distinguées en trois catégories :

1. **Informations provenant de l'entrée vocale :**
   - La *transcription* de la parole, c'est-à-dire la suite des mots de l'énoncé oral.
   - L'*intensité* moyenne de chaque mot (ce qui permet par exemple de savoir si le mot *"cette"* a été accentué dans l'expression *"dans cette boîte"*, afin de prévoir de manière plus sûre la présence d'un geste).
   - Les *dates* de début et de fin de chaque mot, afin d'analyser la synchronisation entre gestes et expressions référentielles.

   Dans un système de dialogue, ces informations sont données par le module de reconnaissance vocale. Compte tenu des performances des systèmes de reconnaissance, plusieurs hypothèses pondérées de cet ensemble d'informations seront données et utilisées dans notre traitement.

2. **Informations provenant de l'entrée gestuelle :**
   - La *liste des objets candidats* (éventuellement vide) pour chaque geste de désignation. Chaque objet sera identifié par un *pointeur* sur l'objet correspondant de l'application, ce qui permettra de retrouver toutes ses caractéristiques. En tenant compte des ambiguïtés de forme et des ambiguïtés de portée, plusieurs hypothèses peuvent être proposées pour chaque geste.
   - Les *dates* de début et de fin de chaque geste, afin d'analyser la synchronisation entre gestes et expressions référentielles.

   Dans un système de dialogue, ces informations sont données par le module de reconnaissance et d'interprétation contextuelle des gestes de désignation.

3. **Informations provenant de l'analyseur syntaxique**, pour chaque hypothèse de l'entrée vocale :
   - L'*arbre de dérivation* obtenu en sortie de l'analyseur syntaxique, en particulier les groupes nominaux sous la forme d'une tête et de modifieurs.
   - Éventuellement, la présence d'une ou de plusieurs *réparations liées à la langue parlée* (comme nous l'avons vu en 1.1.4 : répétition, correction, etc).

Encore une fois, plusieurs hypothèses peuvent être proposées. Dans l'exemple (15), deux hypothèses seront proposées selon le type de réparation syntaxique : l'analyseur syntaxique peut interpréter l'énoncé comme une correction (et le groupe nominal traité ensuite sera *"[les] formes grises à petits pois"*) ou comme une précision (et le groupe nominal traité ensuite sera *"les deux objets/formes grises à petits pois"*). La deuxième hypothèse contient une information supplémentaire très importante : le nombre d'objets.

(15) *"Les deux objets euhhh formes grises à petits pois... dans la troisième boîte"*

Ces informations seront toutes potentiellement disponibles dans un éventuel futur système de dialogue construit à partir des travaux de [Lopez 99] et de [Wolff 99], et utilisant un système de reconnaissance automatique de la parole tel que ViaVoice d'IBM. Par contre, dans notre cadre d'étude du corpus MagnétOz, aucune de ces informations n'est véritablement disponible :

- Les informations liées à l'entrée vocale doivent toutes être reconstituées à partir de l'enregistrement *audio* faisant partie du corpus. Ceci s'avère extrêmement difficile car, lors de la constitution de ce corpus, le but du magicien était le recueil de gestes et non d'expressions référentielles. Ainsi, la qualité la plus faible de fichier *audio* a été choisie. En conséquence, la transcription a été faite à la main (lors de la constitution d'un fichier XML regroupant toutes les informations du corpus), et ni la prosodie ni les dates de début et de fin de mots n'ont été calculées et codées dans ce fichier XML.
- Les informations liées à l'entrée gestuelle ont été spécifiées ici d'une manière différente de celle de Frédéric Wolff (qui ne propose qu'une hypothèse par type d'accès, c'est-à-dire une hypothèse pour un accès individuel et une hypothèse pour un accès pluriel, et qui ne tient donc pas compte de l'ambiguïté de portée).
- Les informations liées à l'analyse syntaxique doivent être générées à partir du corpus lui-même : le lexique du corpus doit être créé et mis sous une forme standard (fichier XML regroupant les lemmes) ; les structures possibles d'énoncés doivent être identifiées. Ce travail n'a pas pu être fait dans les temps.

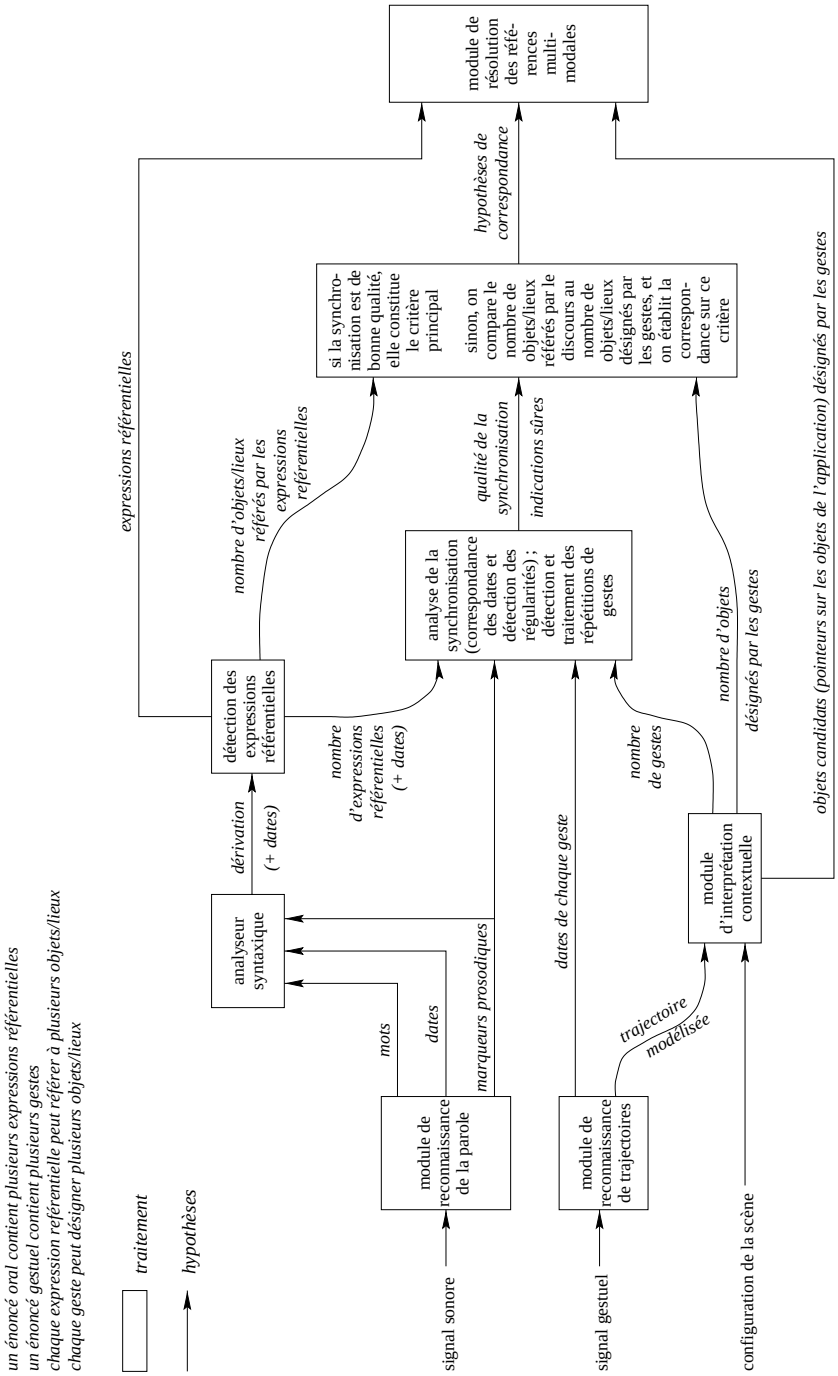
C'est pour cette raison que nous proposons un traitement qui n'a encore été ni implanté ni testé. Notons que, si cela avait été le cas, l'algorithme n'aurait pu être testé que sur le corpus MagnétOz, c'est-à-dire le corpus à l'origine des spécifications de l'algorithme. En effet :

– Il n'existe actuellement aucun autre corpus multimodal.
– Faire des tests sur des utilisateurs nécessite l'implantation d'au moins une partie d'un système de dialogue, ce qui n'est pas envisageable dans le cadre d'un stage de DEA.

Notons cependant que les énoncés multimodaux dont nous tenons compte sont d'une complexité élevée. Notre traitement est donc a priori valable pour tout type d'application permettant de manipuler des objets et des lieux. C'est le cas de toutes les applications citées dans l'introduction de la première partie.

### 4.1.3 Vers un algorithme de traitement

Le but est de trouver le meilleur paramètre possible pour faire les correspondances entre gestes et expressions référentielles. Dans la version actuelle de l'algorithme de traitement présenté dans cette section (se reporter au schéma de la page suivante), deux paramètres vont ainsi être testés l'un après l'autre :




- *Les dates de début et de fin de chaque geste et de chaque expression référentielle :* à l'aide de comparaisons de ces dates, le système va déterminer si certains gestes (voire tous les gestes) ont été effectués en même temps que certaines expressions référentielles. Si ce cas se présente, ce qui arrive souvent dans le corpus, la correspondance est figée pour le geste et l'expression référentielle en question. À cette correspondance est associé un score de qualité de la synchronisation. Ainsi, si le module de résolution des références multimodales aboutit à une erreur et demande une nouvelle hypothèse de correspondance à notre module,

celui-ci rendra son calcul de score plus souple, ce qui permettra de proposer une nouvelle hypothèse. La prosodie intervient en particulier dans le calcul de score.

- *Le nombre d'objets désignés par tous les gestes de l'énoncé et le nombre d'objets référés par toutes les expressions référentielles (s'il est précisé ou s'il se déduit de l'énoncé oral) :* à l'aide de comparaisons, le système va essayer de trouver une équivalence des cardinaux (en partant du premier geste et en y ajoutant les autres gestes un par un). Généralement, une et une seule hypothèse de candidats aux gestes de désignation a le même cardinal que celui obtenu des expressions référentielles.

D'autre part, le traitement tient compte des répétitions et des corrections de la manière suivante : si l'énoncé oral contient la correction d'une expression référentielle et si un geste est associé à chacune des deux expressions, alors le premier geste est négligé de la même façon que la première expression.

Enfin, un geste ne pouvant correspondre à aucune expression référentielle sera considéré (du moins dans certaines des hypothèses) comme un geste involontaire.

## 4.2 Synthèse théorique de l'analyse

### 4.2.1 Point de départ : informations explicites et informations implicites

Selon une première approche de résolution des références, on suppose que toutes les connaissances nécessaires à la résolution sont contenues dans l'énoncé, c'est-à-dire qu'elles sont explicites. Or ce n'est pas toujours le cas dans le dialogue finalisé, du fait du caractère très orienté sur la tâche à accomplir. Autrement dit, la conscience de la tâche entraîne l'utilisateur à faire des sous-entendus. Ces sous-entendus (ou informations implicites) sont sensés être retrouvés facilement par l'interlocuteur. Lorsqu'il s'agit d'une machine, retrouver ces informations s'avère en fait un problème très complexe.

Par exemple, un énoncé tel que : *"peins la chaise"* (sans geste) peut correspondre à des actions très diverses selon l'information implicite qui peut lui être associée. La liste suivante montre la variété de l'implicite :

| énoncé | implicite sur l'objet | implicite sur la couleur |
|---|---|---|
| *"peins la chaise"* | [celle que je viens de manipuler]<br>[celle à laquelle je n'ai pas encore touché]<br>[la seule à l'écran]<br>[la seule saillante]<br>[la seule à ne pas être peinte]<br>[celle qui a l'air d'une antiquité]<br>[celle que je montre du doigt]<br>[celle que je regarde]<br>[celle juste à côté du meuble que je viens de manipuler]<br>[n'importe laquelle, mais fais quelque chose au lieu de refuser tous mes énoncés !]<br>[... enfin le fauteuil, pardon !] | [de la couleur avec laquelle je viens de peindre une autre chaise]<br>[de la couleur utilisée précédemment pour peindre la table]<br>[de la seule couleur qu'ont les autres meubles de la pièce, sauf justement cette chaise]<br>[de la seule couleur que j'ai utilisée jusqu'à présent]<br>[de n'importe quelle couleur, mais fais quelque chose au lieu de refuser tous mes énoncés !]<br>etc... |

La plupart des informations implicites possibles sont fortement liées à la tâche : ainsi, celles

présentées dans le tableau ci-dessus portent fréquemment sur les actions précédentes ou sur la perception visuelle. Le dialogue homme-machine met en jeu beaucoup d'implicite et il est nécessaire de retrouver les informations implicites pour comprendre les énoncés. L'exemple le plus flagrant dans le corpus est lié au fait de ranger et d'avoir rangé des objets : dans une scène comportant trois objets dont l'un est visiblement rangé, l'utilisateur va dire *"range les deux objets"* sans faire de geste. L'information implicite (ou pré-condition dans le cas de cet exemple) à retrouver est *"range les deux objets [qui ne sont pas encore rangés]"*. Si l'on ne tient pas compte de cette information, l'énoncé est incompréhensible car le nombre d'*objets* présents dans la scène n'est pas celui indiqué dans l'énoncé. Par contre, si l'on tient compte de cette information, le nombre d'*objets qui ne sont pas encore rangés* correspond bien à celui indiqué dans l'énoncé. Autant l'information implicite est ici facile à retrouver grâce au prédicat, autant ce n'est pas toujours le cas.

La saillance est un phénomène qui rend un ou plusieurs objets susceptibles *plus que les autres* de faire intervenir des informations implicites. Ainsi, lorsque par exemple au théâtre un acteur en train de jouer est éclairé par les projecteurs, il est fortement saillant et peut être désigné (par un spectateur à son voisin) par la simple expression *"l'acteur"* ou *"il"*, même si d'autres acteurs sont juste à côté de lui. Dans le cadre d'une application de dialogue finalisé, les objets peuvent être rendus volontairement saillants (par exemple en les faisant clignoter), ou peuvent être saillants du fait de leur état ou de leur situation dans la scène. Dans ce cas, l'application elle-même ne sait pas quels objets sont saillants. C'est pourtant un paramètre important lors de l'interprétation des énoncés. Il est donc utile voire nécessaire d'identifier les paramètres pouvant rendre un objet saillant. Ce problème est étudié dans [Landragin 98] qui aboutit à la classification des paramètres suivants :

1. Saillance par la **catégorie** de l'objet : dans une scène contenant plusieurs chaises et une table, la table est saillante.
2. Saillance par ses **caractéristiques physiques** (taille, géométrie, matériau, couleur, texture, etc) : dans une scène contenant des chaises dont une qui est plus petite que les autres, cette dernière est saillante.
3. Saillance par ses **fonctionnalités** : dans une scène contenant plusieurs ordinateurs dont un en fonctionnement, ce dernier est saillant (les fonctionnalités peuvent être perçues visuellement).
4. Saillance **spatiale** ou saillance par sa **localisation dans la scène** (par rapport aux autres objets et par rapport à l'utilisateur) : dans une scène en *3D* contenant des chaises, une chaise très proche est saillante si les autres sont éloignées ; une chaise isolée est saillante si les autres sont groupées.
5. Saillance par son **incongruité** (lorsque l'objet est en infraction avec une règle implicite, culturelle ou fonctionnelle) : dans une scène contenant des chaises dont une qui est renversée à même le sol, cette dernière est saillante.
6. Saillance par sa **dynamique** : dans une scène contenant un objet en mouvement et d'autres statiques, l'objet en mouvement est saillant.

### 4.2.2 Nombre et validité des informations données dans l'énoncé oral

Dans le corpus MagnétOz, certaines expressions référentielles accumulent les informations décrivant les objets référencés, alors qu'un geste de désignation est produit de manière précise et non ambiguë. Lors de la compréhension, le filtre perceptif suffit à déterminer les référents et

toutes les informations données dans l'énoncé oral peuvent donc être négligées. C'est d'ailleurs ce que fait le magicien.

Il est intéressant cependant d'étudier de manière approfondie ces informations. On remarque que dans la majorité des cas, elles sont inutiles et ne font que confirmer le geste. Nous sommes alors dans la relation de *redondance* entre modalités, et le filtre perceptif seul est suffisant. Par contre, dans deux cas particuliers du corpus, ces informations vont à l'encontre du geste, c'est-à-dire que nous sommes alors dans une relation de *contradiction* entre modalités : le filtre perceptif et le filtre linguistique donnent des résultats différents, dont l'intersection est vide. Cette constatation entraîne les questions suivantes :

- Pourquoi l'utilisateur donne-t-il ces informations alors que le geste n'est pas ambigu et suffit à déterminer les référents ?
- L'utilisateur s'est-il trompé en voulant donner trop d'informations, ou, au contraire, a-t-il agit selon un mécanisme utilisé également dans ses autres énoncés, mécanisme que nous n'avons pas identifié jusqu'à maintenant ?
- Comment traiter les informations linguistiques lorsque le geste n'est pas ambigu ?

Afin de proposer des réponses à ces questions, nous allons considérer les deux cas particuliers du corpus :

(16) *"l'objet le plus à droite"* + geste sur un objet qui n'est pas le plus à droite dans la scène (dans le coin supérieur droit de la scène se trouve un autre objet qui, lui, vérifie les conditions de l'énoncé oral seul)

(17) *"les formes les plus claires"* + geste sur trois formes claires formant un groupe, entourées de formes sombres, une forme très claire se trouvant plus loin dans un coin de la scène.

Nous remarquons dans les deux cas que, si l'on tient compte de la scène complète, les propriétés données dans l'énoncé oral s'appliquent à d'autre objets que ceux intentionnellement désignés. Le filtre linguistique ne nous est donc d'aucune utilité ici. Si d'autre part on ne tient compte que de l'espace perceptif délimité par le geste, les propriétés données dans l'énoncé n'ont aucun sens puisqu'il s'agit de superlatifs et que l'espace perceptif ne contient que des objets identiques. Pour que ces propriétés aient un sens, il faut étendre l'espace perceptif au voisinage immédiat du geste, et c'est probablement ce qu'a fait l'utilisateur lorsqu'il a choisi les termes de son énoncé. On peut trouver la raison de cette extension dans la volonté d'exprimer un contraste entre formes claires et formes sombres, ce contraste étant saillant visuellement et faisant partie du mécanisme de production et de compréhension des expressions définies (comme le montre le schéma axiologique du défini).

Le principe consistant à étendre un geste même quand ce geste est effectué de manière précise n'est pas aberrant. En plus de la possibilité que la précision ne soit qu'un hasard, nous noterons que ce n'est pas parce qu'un geste délimite précisément une zone que le regard de l'utilisateur s'arrête aux limites de cette zone. Comme nous venons de le voir, la zone peut être étendue jusqu'à contenir un contraste. Elle peut aussi être étendue jusqu'à contenir un objet facilement nominalisable (en supposant que ce n'est pas le cas dans la zone), cet objet étant alors celui nommé dans l'énoncé. Or, si le regard de l'utilisateur ne s'arrête pas aux limites de la zone désignée par le geste, il est logique de dire que le regard de l'interlocuteur ne le fait pas non plus, et qu'un système ne doit pas se limiter à un filtre perceptif figé.

Nous avons maintenant une réponse possible pour chacune des questions posées :

- L'utilisateur donne des informations linguistiques pour exprimer un contraste afin de justifier son emploi d'une expression définie.

- Il ne s'est pas trompé mais agit selon un mécanisme que nous n'avions pas identifié jusqu'à maintenant (le contraste étant rarement explicite dans l'énoncé).
- Ces informations ne doivent pas être négligées : au contraire, l'extension de l'espace perceptif doit être effectuée afin de vérifier la présence du contraste et de le comprendre.

D'autre part, un rapport entre le nombre d'informations données dans l'énoncé et l'importance de l'implicite peut être considéré : en supposant que la somme des informations explicites et des informations implicites reste stable autour d'une certaine moyenne pour chaque type d'énoncé, on déduit (du point de vue de la compréhension) que plus le nombre d'informations données est faible, plus l'implicite est important (et, inversement, que plus le nombre d'informations données est important, plus l'implicite est négligeable). Du point de vue de la production des énoncés, on déduit que plus l'utilisateur aura conscience de la saillance de l'implicite, plus il va réduire son énoncé, et, au contraire, plus l'implicite sera mal saillant, plus l'utilisateur va donner d'informations.

### 4.2.3 Expression définie et expression démonstrative

Dans le corpus MagnétOz, le geste apparaît aussi bien avec des expressions définies qu'avec des expressions démonstratives, sans qu'une quelconque différence soit faite entre les deux au niveau de la compréhension de l'énoncé multimodal : les gestes sont les mêmes, la production (rythme, intonation) des énoncés est comparable. Dans de nombreux cas, le geste apparaît aussi avec un défini associé à un marqueur déictique *("le N-ci" ou "le N-là")*, qui est une forme grammaticalement incorrecte. On peut alors se poser les questions suivantes :

- La présence d'un geste ne donnerait-elle pas au défini un certain caractère démonstratif, les deux types de déterminant obéissant alors à un même mécanisme ?
- Indépendamment de la question précédente, à quel mécanisme obéit l'association d'un défini et d'un marqueur déictique ?
- La méthode classique consistant à prévoir l'apparition d'un geste avec un démonstratif mais pas avec un défini est-elle toujours valable ?

Nous avons vu en 2.2.2 que, dans le discours seul, le défini et le démonstratif suivaient deux mécanismes distincts. Il est nécessaire de revenir sur ces mécanismes et de faire le lien avec le traitement classique de la multimodalité (filtre perceptif et filtre linguistique).

Lorsque l'interlocuteur entend un groupe nominal défini comme par exemple *"le triangle"* dans une scène contenant des objets géométriques, il va partir d'un domaine de référence contenant des triangles et des objets qui ne sont pas des triangles, puis faire l'opposition (axiologie) entre les triangles et les non-triangles. Autrement dit, il va partir du domaine de référence pour arriver à l'objet. Lorsqu'un geste accompagne l'expression, le domaine de référence va être construit immédiatement puisque c'est le sous-espace perceptif délimité par le geste qui va le déterminer. Le domaine de référence pourra être étendu au voisinage immédiat du geste comme nous l'avons vu en 4.2.2. L'interlocuteur va donc d'abord percevoir tous les objets contenus dans ce domaine de référence, percevoir l'axiologie, puis ne considérer que les objets de la catégorie triangle pour retrouver le référent. Autrement dit, il va faire le filtre perceptif avant le filtre linguistique.

Par contre, lorsque l'interlocuteur entend un groupe nominal démonstratif comme par exemple *"ce triangle"* accompagné d'un geste de désignation, il va partir de l'objet de la catégorie triangle pour arriver au domaine de référence qui contiendra le triangle désigné et les autres triangles de la scène. Autrement dit, il va faire le filtre linguistique avant le filtre perceptif.

D'une manière générale, les schémas axiologiques sont particulièrement intéressants pour préparer les reprises. L'expression *"le triangle"* pourra être suivie de *"puis le carré"* mais sans doute pas de *"puis l'autre"*. Par contre, l'expression *"ce triangle"* pourra être suivie de *"puis l'autre"* mais sans doute pas de *"puis le carré"*. Ce sont surtout les mécanismes de reprise qui font la différence entre défini et démonstratif.

Deux mécanismes distincts sont donc bien en jeu. Le corpus MagnétOz ne permet pas de les distinguer facilement pour la simple raison que les reprises y sont quasiment inexistantes. La différence entre défini et démonstratif peut se retrouver par l'implicite de la manière suivante :

(18) *"le triangle"* + geste = *"un objet que je désigne par ce geste et qui est de la catégorie triangle"* + geste.

(19) *"ce triangle"* + geste = *"le triangle que je désigne par ce geste"* + geste.

D'autre part, un groupe nominal défini comportant un marqueur déictique tel que *"la boîte-ci"* ne peut pas vraiment être repris par *"puis le coffre"* ni même par *"puis l'autre"* (l'énoncé paraît incomplet). A priori, ce type de groupe nominal ne suit ni le mécanisme du défini, ni celui du démonstratif. La remarque de l'énoncé qui paraît incomplet permet de trouver une réponse : *"la boîte-ci"* peut être repris par *"puis l'autre là"*. Or ceci n'a qu'une explication possible qui apparaît lorsque l'énoncé est transcrit de la manière suivante : *"la boîte, /i/ci, puis l'autre, là"*, c'est-à-dire avec des déictiques purs indépendants des groupes nominaux et non des marqueurs déictiques liés aux groupes nominaux. La transformation de *"ici"* en *"ci"* peut s'expliquer à l'aide du phénomène de coarticulation vu en 1.1.2. Notre hypothèse est d'une certaine manière confirmée par la présence dans le corpus d'expressions telles que *"la boîte ici"* (du point de vue théorique de la linguistique diachronique, les deux formes peuvent cohabiter).

Nous avons maintenant une réponse possible pour chacune des questions posées :

- Défini et démonstratif obéissent à deux mécanismes différents, même lorsqu'ils sont accompagnés par un geste.
- Ce que nous avons pris pour l'association d'un défini et d'un marqueur déictique peut en fait correspondre à un groupe nominal défini suivi d'un déictique pur. La référence en jeu est donc celle du déictique pur.
- Un geste pouvant être effectué aussi bien avec un défini qu'avec un démonstratif, les apparitions de geste sont impossibles à prévoir sur ce seul critère.

### 4.2.4 Mécanismes de désignation

Dans cette section, nous reprenons la dernière partie de l'état de l'art sur la multimodalité (section 1.3.2), et nous y apportons les résultats de notre analyse ainsi que les résultats de notre réflexion théorique suite à cette analyse.

Une référence multimodale est l'association d'une expression référentielle et d'un geste. Son résultat est l'objet ou le lieu co-référé par les deux modalités. Selon la fonction de l'expression référentielle dans l'énoncé, l'intention de désignation peut varier : l'utilisateur peut par exemple vouloir désigner une couleur et faire une référence à un objet ayant cette couleur. Ainsi, une référence se fait sur un objet ou un lieu, et le résultat de la référence est l'objet ou le lieu en question ; une désignation se fait également sur un objet ou un lieu, mais son résultat :

– peut être un objet, un lieu, ou une caractéristique d'un objet,
– peut être différent du résultat de la référence (nous parlerons alors de désignation indirecte).

À partir des trois paramètres que sont objet, caractéristique et lieu, nous proposons une classification des mécanismes de désignation :

- **Désignation directe d'objets :** lorsqu'un objet est un argument de l'action à effectuer, l'utilisateur peut décrire l'objet dans l'énoncé oral et/ou le désigner par un geste *(accès individuel)* ; lorsque plusieurs objets sont des arguments de l'action, l'utilisateur peut soit désigner chaque objet l'un après l'autre *(accès multiple, c'est-à-dire série d'accès individuels)*, soit désigner les objets par un seul geste s'ils forment un groupe perceptif au sens de la *proximité (accès pluriel)*. Si les objets forment un groupe perceptif au sens de la *similarité*, l'accès multiple sera probablement utilisé.
  Les références multimodales combinées que nous avons étudiées sont des combinaisons de ces trois types d'accès.

  (20) accès individuel : *"cet objet"* + un geste sur un objet.
  (21) accès multiple : *"cet objet, cet objet et celui-ci"* + trois gestes, chacun sur un objet.
  (22) accès pluriel : *"ces objets"* + un geste sur trois objets à la fois.

- **Désignation directe de lieux :** comme pour la désignation directe d'objets, l'utilisateur peut décrire le lieu dans l'énoncé oral et/ou le désigner par un geste ; l'accès multiple de lieux et l'accès pluriel sont également possibles, bien que peu fréquents. Notons également que la désignation d'un lieu pose le problème de l'étendue de ce lieu : dans l'exemple (26) [Romary 93] et dans l'exemple (27), le même geste désigne des lieux de tailles très différentes. Ce problème rejoint celui de l'ambiguïté de portée du geste identifiée dans les désignations de groupes perceptifs d'objets.

  (23) accès individuel : *"ici"* + un geste sur un lieu.
  (24) accès multiple : *"ici et là"* + deux gestes, chacun sur un lieu.
  (25) accès pluriel : *"partout là"* + un geste sur plusieurs lieux ou sur une étendue spatiale.
  (26) *"mettre de la moquette ici"* + un geste de pointage.
  (27) *"planter un clou ici"* + un geste de pointage.

- **Désignation indirecte de caractéristiques :** lorsqu'une caractéristique est un argument de l'action, cette caractéristique peut être désignée à l'aide d'un geste sur un objet qui la possède. Dans l'exemple (28), la caractéristique de couleur est *explicite*. Plusieurs caractéristiques peuvent être associées : en reprenant l'exemple précédent, les caractéristiques retenues peuvent être non seulement la couleur, mais aussi la brillance et la texture, qui sont probablement *implicites*. Toutes les caractéristiques peuvent d'ailleurs être *implicites*, comme dans l'exemple (29). Dans cet exemple, le prédicat permet de retrouver la caractéristique principale. D'autre part, dans l'exemple (30), c'est le mécanisme de désignation indirecte lui-même qui est implicite (il ne s'agit pas de désignation directe de couleur car la couleur est une caractéristique immatérielle, autrement dit la *désignation directe de caractéristiques* n'existe pas).

  (28) caractéristique explicite : *"peindre la table de la couleur de ces chaises"* + un geste sur un groupe de chaises.
  (29) caractéristique implicite : *"peindre la table comme ça"* + un geste sur un groupe de chaises.
  (30) mécanisme de désignation indirecte implicite : *"peindre la table de cette couleur"* + un geste sur une chaise.

Dans tous ces exemples, le mécanime de désignation est le suivant : l'utilisateur part d'un *objet* pour en extraire une *caractéristique*.

- **Désignation indirecte d'objets :** un ensemble d'objets ne formant pas de groupe perceptif (au sens de la proximité) peut être désigné autrement que par une désignation multiple. En effet, si les objets ont des caractéristiques similaires, l'utilisateur peut utiliser ce critère et effectuer une *désignation indirecte de caractéristiques* sur *un* objet de l'ensemble. La ou les caractéristiques peuvent être *explicites* comme dans les exemples (31) et (32), ou *implicites* comme dans les exemples (33) et (34). L'utilisateur peut même rendre *implicite* le mécanisme de désignation indirecte, comme dans l'exemple (35) équivalent aux exemples (33) et (34).

  (31) caractéristique explicite retrouvée par un mécanisme de désignation indirecte explicite : *"les fauteuils de la couleur de celui-ci"* + un geste sur un fauteuil ; *"le fauteuil à côté de celui-ci"* + un geste sur un fauteuil.

  (32) caractéristique explicite retrouvée par un mécanisme de désignation indirecte implicite : *"les fauteuils de cette couleur"* + un geste sur un fauteuil.

  (33) caractéristique implicite retrouvée par un mécanisme de désignation indirecte explicite : *"les fauteuils du genre de celui-ci"* + un geste sur un fauteuil ; *"les fauteuils comme celui-ci"* + un geste sur un fauteuil.

  (34) caractéristique implicite retrouvée par un mécanisme de désignation indirecte implicite : *"les fauteuils de ce genre"* + un geste sur un fauteuil.

  (35) mécanisme de désignation indirecte implicite : *"j'aime bien ces fauteuils"* + un geste sur un fauteuil.

  Dans tous ces exemples, le mécanime de désignation est le suivant : l'utilisateur part d'un *objet* pour en extraire une *caractéristique* et construire à partir de cette caractéristique un objet ou un ensemble d'*objets*.

- **Désignation indirecte de lieux :** l'utilisateur peut désigner un lieu en effectuant un geste sur un objet placé dans ce lieu. Ici aussi, le mécanisme de désignation indirecte peut être *explicite* comme dans l'exemple (36), ou *implicite* comme dans l'exemple (37).

  (36) mécanisme de désignation indirecte explicite : *"ajouter une chaise à côté de celle-ci"* + un geste sur une chaise.

  (37) mécanisme de désignation indirecte implicite : *"ici"* + un geste sur un objet.

  Dans ces exemples, le mécanime de désignation est le suivant : l'utilisateur part d'un *objet* pour désigner le *lieu* dans lequel se trouve l'objet.
  Nous remarquons ici que le mécanisme consistant à partir d'un *lieu* pour désigner le ou les *objets* se trouvant en ce lieu n'a pas été explicité. Avec ce que nous avons vu en 4.2.3, les expressions associant groupe nominal défini et déictique pur en sont des exemples typiques, comme : *"la boîte là"* + geste. En fait, un problème théorique apparaît, sur la distinction que l'on fait entre objet et lieu : comme la désignation d'objet suit à peu près les mêmes mécanismes que la désignation de lieu, comme les gestes effectués dans les deux cas sont aussi variés, comme l'ambiguïté entre désignation d'objet et désignation de lieu est fréquente, on peut se demander si objet et lieu ne peuvent pas être considérés comme un même concept sur lequel baser les mécanismes de désignation. Cette question reste ouverte...

Les exemples que nous proposons dans cette classification montrent la complexité de la compréhension d'un énoncé multimodal. Une interprétation correcte nécessite en effet de retrouver à partir des informations données par les énoncés oral et gestuel le mécanisme de désignation, ce qui peut s'avérer complexe. Si le système ne retrouve pas le mécanisme de désignation, il peut aboutir à une erreur. Ainsi, dans l'exemple (35), les informations semblent a priori se contredire : l'expression référentielle est au pluriel alors que le geste ne désigne qu'un objet. Cette contradiction résulte en fait d'un mécanisme complexe partant d'un objet pour en extraire une ou plusieurs caractéristiques implicites à partir desquelles est construit un ensemble d'objets. Dans la vie courante, ce mécanisme est utilisé tous les jours, sans que nous ayons conscience de sa complexité.

# Conclusion et perspectives

**Récapitulatif sur la compréhension d'un énoncé multimodal**

La compréhension d'un énoncé multimodal associant la parole et le geste de désignation repose sur de nombreux mécanismes : modes de désignation *(désignation directe d'un objet, désignation indirecte d'une caractéristique d'un objet...)*, modes de référence *(référence démonstrative, référence déictique...)*, structures du discours parlé *(présence d'ellipses, de répétitions, de corrections...)*. Ces mécanismes complémentaires se combinent pour fournir une interprétation qui, dans le cadre d'un dialogue de commande, doit conduire à la réalisation correcte de l'intention que l'utilisateur a cherché à exprimer.

Deux catégories de désignation utilisant la parole et le geste sont classiquement distinguées : la désignation d'objet(s) et la désignation de lieu(x). Les modalités concourent dans les deux cas à l'identification d'une même entité. Ce phénomène est appelé co-référence ou référence multimodale. L'interprétation des informations données par les modalités se fait par une combinaison de filtres : un filtre linguistique permet de ne considérer que les objets vérifiant les conditions de l'énoncé oral, et un filtre perceptif réduit l'espace de recherche des référents au sous-espace perceptif désigné par le geste. Cette méthode simpliste d'analyse de l'articulation entre parole et geste s'avère réductrice face à la diversité des énoncés multimodaux spontanés. Le besoin d'une méthode d'analyse véritablement multimodale se fait alors ressentir. De plus, la désignation peut prendre d'autres formes que les deux citées : elle peut d'une part extraire de l'environnement non seulement un objet ou un lieu, mais aussi une caractéristique d'un objet, d'autre part extraire un objet ou un lieu différent de celui référé. Un système de dialogue doit être capable de retrouver l'intention multimodale de l'utilisateur, c'est-à-dire en particulier le mode de désignation qu'il a adopté et le type de référence multimodale qu'il a utilisé.

La possibilité de désigner des objets par le geste permet à l'utilisateur de réduire la description de ces objets dans l'énoncé oral. De plus, dans le cadre du dialogue finalisé, l'importance de la tâche à effectuer entraîne l'utilisateur à faire des sous-entendus, qui sont sensés être compris facilement dans le contexte applicatif. D'une manière générale, il faut considérer qu'il existe des informations implicites liées à l'énoncé multimodal. L'interprétation de ces informations dépend du geste effectué ainsi que des contextes applicatif, perceptif et discursif. Un système de dialogue doit les retrouver pour comprendre l'intention de l'utilisateur.

- Une première façon de retrouver l'implicite à l'aide du *contexte applicatif* est de tenir compte des pré-conditions : dans l'exemple *"range la chaise"*, le système ne doit chercher une chaise que dans l'ensemble des chaises qui ne sont pas encore rangées.

- Une deuxième façon de retrouver l'implicite à l'aide du *contexte perceptif* est de tenir compte de la saillance perceptive des objets : un objet saillant possède une caractéristique

particulière (due à son état ou à sa situation dans l'environnement) qui le met en valeur par rapport aux autres objets, et qui fait intervenir fortement l'implicite lors de sa désignation. En reprenant l'exemple *"range la chaise"*, si plusieurs chaises sont des candidats à la référence, l'une d'entre elles peut être saillante et correspondre à l'intention de l'utilisateur.

- Une troisième façon de retrouver l'implicite à l'aide du *contexte discursif* est de considérer l'anaphore lorsque l'historique du dialogue contient une chaise récemment manipulée. L'exemple *"range la chaise"* peut également être interprété de cette dernière façon, ceci montrant l'interférence possible entre contexte perceptif et contexte discursif.

**Travail effectué et perspectives**

À partir de l'analyse d'un corpus multimodal, nous avons identifié et décrit des mécanismes de désignation, ainsi que les problèmes de compréhension qui leur sont liés, en particulier au niveau de la référence. Nous avons d'autre part identifié le problème technique consistant à établir les correspondances entre gestes et expressions référentielles dans les énoncés faisant intervenir plusieurs références multimodales ou des références multimodales combinées (phénomène que nous avons identifié, correspondant à la production de $p$ gestes avec $q$ expressions référentielles). Nous avons également proposé une première modélisation du traitement de ces énoncés.

Les prolongements possibles de l'étude sont, comme notre démarche, d'ordre théorique et d'ordre technique : à partir d'études théoriques sur l'articulation entre parole et geste dans la communication, et de considérations techniques, le but est d'aboutir à la modélisation d'un système de dialogue capable de comprendre les intentions multimodales de l'utilisateur. Plus particulièrement, il s'agit de :

- étudier les mécanismes intervenant lors de la production des énoncés multimodaux *(choix du mode de désignation, du mode de référence, du déterminant et des informations explicites compte tenu de l'implicite)* ; identifier de manière approfondie les problèmes qui se posent lors de la compréhension de ces énoncés *(en particulier les problèmes d'ordre pragmatique)* ; étudier les possibilités d'adaptation du modèle des *Représentations Mentales* des référents à la multimodalité *(vers un concept unifié du traitement de la référence)* ;

- proposer une modélisation *(en tenant compte des différents modes de désignation, des différents modes de référence, et de la variété des énoncés multimodaux pour chacun de ces modes)* et une formalisation *(le but étant d'aboutir à un résultat pouvant être intégré dans une architecture de résolution des références)* ;

- réaliser une implantation *(à laquelle il sera possible d'ajouter la gestion d'un historique de la perception en rapport avec l'historique du discours, ainsi que la gestion de profils utilisateurs contenant en particulier leurs préférences d'utilisation de la multimodalité)* ; valider cette implantation sur des corpus ou sur des systèmes de dialogue existants.

# Bibliographie


[Bellalem 95] N. Bellalem. *Étude du mode de désignation dans un dialogue homme-machine finalisé à forte composante langagière : analyse structurelle et interprétation.* Thèse de Doctorat, Université de Nancy 1, 1995.

[Braffort 96] A. Braffort. *Reconnaissance et compréhension de gestes, application à la langue des signes.* Thèse de Doctorat, Université de Paris 11, 1996.

[Brison 97] E. Brison. *Stratégies de compréhension dans l'interaction multimodale.* Thèse de Doctorat, Université Paul Sabatier de Toulouse, 1997.

[Cadoz 94] C. Cadoz. Le geste canal de communication homme/machine – la communication instrumentale. *Technique et Sciences Informatiques*, 13(1):31–61, 1994.

[Carré 91] R. Carré, J.-F. Dégremont, M. Gross, J.-M. Pierrel et G. Sabah. *Langage humain et machine.* Presses du CNRS, Paris, 1991.

[Coutaz 91] J. Coutaz et J. Caelen. A taxonomy for multimedia and multimodal user interfaces. *ERCIM Workshop, INESC*, pages 143–148, Lisbon, Portugal, 1991.

[Duermael 94] F. Duermael. *Référence aux actions dans des dialogues de commande homme-machine.* Thèse de Doctorat, Institut National Polytechnique de Lorraine, 1994.

[Gaiffe 92] B. Gaiffe. *Référence et dialogue homme-machine : vers un modèle adapté au multi-modal.* Thèse de Doctorat, Université de Nancy 1, 1992.

[Gaiffe 94] B. Gaiffe, A. Reboul et L. Romary. Références et gestion du dialogue. *Actes de TALN'94*, Marseille, 1994.

[Guyomard 95] M. Guyomard, D. Le Meur, S. Poignonnec et J. Siroux. Experimental work for the dual usage of voice and touch screen for a cartographic application. *ESCA Workshop on Spoken Dialogue Systems*, pages 153–156, 1995.

[Husson 98] J.-L. Husson. *Une approche hiérarchique de la segmentation du signal de parole.* Thèse de Doctorat, Université Henri Poincaré, Nancy, 1998.

[Kamp 88] H. Kamp. Discourse representation theory, what it is and where it ought to go. *Scientific Symposium on Syntax and Semantics for Text Processing and Man Machine Communication*, Heidelberg, 1988.

[Kerbrat-Orecchioni 96] C. Kerbrat-Orecchioni. *La conversation.* Seuil, 1996.

[Lambrecht 96] K. Lambrecht. On the formal and functional relationship between topics and vocatives. *Conceptual Structure, Discourse and Language, Stanford, CSLI Publications*, pages 267–288, 1996.

[Landragin 98] F. Landragin. Interaction multimodale dans un environnement virtuel. rapport de stage de fin d'études de l'Institut d'Informatique d'Entreprise ; Laboratoire Central de Recherches de THOMSON-CSF, Orsay, 56 pages, 1998.

[Landragin 99] F. Landragin. Rapports entre oral et écrit. rapport interne LORIA, Nancy, 9 pages, 1999.

[Lopez 99] P. Lopez. *Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres.* Thèse de Doctorat, Université de Nancy 1, 1999. à paraître.

[Luzzati 95] D. Luzzati. *Le dialogue verbal homme-machine : étude de cas.* Sciences Cognitives, Masson, 1995.

[Mathieu 97] F.-A. Mathieu. *Prise en compte de contraintes pragmatiques pour guider un système de reconnaissance de la parole : le système COMPPA.* Thèse de Doctorat, Université Henri Poincaré Nancy 1, 1997.

[Mignot 95] C. Mignot. *Usage de la parole et du geste dans les interfaces multimodales. Étude expérimentale et modélisation.* Thèse de Doctorat, Université de Nancy 1, 1995.

[Moulton 94] J. Moulton et L. D. Roberts. An ai module for reference based on perception. P. McKevitt, éditeur, *AAAI workshop on integration of natural language and vision processing*, Seattle, 1994.

[Pierrel 87] J.-M. Pierrel. *Dialogue oral homme-Machine.* Paris, Hermès, 1987.

[Pouteau 94] X. Pouteau, L. Romary et J.-M. Pierrel. Voix, geste et multimodalité : quand dire c'est faire faire. *Actes Congrès ERGO-IA'94*, pages 491–500, Biarritz, 1994.

[Reboul 98] A. Reboul et J. Moeschler. *Pragmatique du discours.* Armand Colin, Paris, 1998.

[Romary 93] L. Romary. "mets ça ici" où quand "ici" dépend de "ça" : l'interprétation de "ici" dans des énoncés de positionnement. *Workshop Le dialogue homme-robot en langage naturel*, Caen, 1993.

[Waibel 95] A. Waibel, M. T. Vo, P. Duchnowski et S. Manke. Multimodal interfaces. *Artificial Intelligence Review, special volume on integration of natural language and vision processing, McKevitt*, 10, 1995.

[Wolff 99] F. Wolff. *Analyse contextuelle des gestes de désignation en dialogue homme-machine.* Thèse de Doctorat, Université de Nancy 1, 1999.

## Résumé

Une communication homme-machine véritablement naturelle se doit d'accepter et de comprendre les moyens naturels que possède l'homme pour communiquer, c'est-à-dire la parole et le geste venant en complément de celle-ci. Dans un système de dialogue homme-machine finalisé, les objets de l'application peuvent être désignés soit par un énoncé oral seul, soit par l'association d'un énoncé oral et d'un geste de désignation, ce qui constitue une interaction multimodale. Ce mémoire présente une étude des associations possibles de la voix et du geste spontanés, effectués sans contrainte, ainsi que de leurs traitements.

Dans l'état de l'art qui constitue la première partie de ce travail, nous nous focalisons sur le problème de la référence aux objets et nous montrons que les méthodes utilisées pour résoudre les références multimodales (c'est-à-dire pour faire le lien entre d'un côté les mots utilisés et les gestes effectués, de l'autre côté les objets de l'application) sont soit des extensions insuffisantes de méthodes utilisées classiquement dans le discours, soit des méthodes trop simplistes pour rendre compte de la complexité des énoncés multimodaux.

Dans la deuxième partie de ce travail, nous étudions un corpus multimodal selon une approche linguistique, complétant ainsi la première analyse de ce corpus effectuée par Frédéric Wolff et centrée sur le geste (thèse de doctorat soutenue en 1999). Nous identifions les informations contenues dans les énoncés oraux et nous les confrontons aux hypothèses de candidats des désignations gestuelles, en tenant compte des contextes discursif, perceptif et applicatif. Apparaît alors le problème consistant à établir les correspondances entre gestes et expressions référentielles lorsque $p$ gestes sont effectués avec $q$ expressions, phénomène formalisé sous le nom de référence multimodale combinée. Les informations temporelles, prosodiques et syntaxiques nécessaires à l'interprétation de ce type de référence sont explicitées, en tenant compte des spécificités de la langue parlée telles que les répétitions ou les auto-corrections pour lesquelles il existe des équivalents dans la production du geste. Une modélisation du traitement de ces informations est alors proposée, puis une synthèse théorique de notre analyse, portant sur la prise en compte d'informations implicites pour la résolution des références multimodales. Cette synthèse permet d'aboutir à la proposition d'une classification des mécanismes de désignation et à des possibilités de prolongements de l'étude.



**Mots-clés:** dialogue homme-machine, interaction multimodale, langue naturelle, désignation, geste, référence